ENGLISH TO TAMIL MACHINE TRANSLATION SYSTEM USING PARALLEL CORPUS

Dr. G.Vasuki
akshayvaasu@gmail.com
&

Prof. Rajendran Sankaravelayuthan
rajushush@gmail.com

Thanjavur
2013

A FEW WORDS

I have been working in the field of MT with special referece to Tamil since 1990. It is a long journey. Still I am working in this area. I got opportunity to work for a few MT projects funded by Govt. of India for the technological development of Indian Languages.

The present work was done by Dr. G. Vasuki for her PhD work under my supervision. It was lying in my lap top since 2013. I thought of editing it and publish as a book. But time is running out. So I thought of uploading it somewhere. With here consent I am presenting it here.

After some time the entire exercise will be irrelevant. A rule based system making use of the theory of linguistics is a wecome thing. But unfortunbately nobody has the patinence for it. Computer scientist want to get of rid of linguistics. They want to try for short cut method. SMT is the outcome of such short-cut method. Google has come out with a comparatively good MT system for English-Tamil translation. The systems built by some of us are inferor to what is available in Google. But a research is a research.

Prof. S. Rajendran

CONTENT	PAGE NOS
Chapter 1 Introduction	9
1.1. Motivation	9
1.2 Issues in the research	10
1.3 Aims and objectives of the work	12
1.4 Methodology	13
1.5. Previous research works	13
1.6 Chaterization	14
1.7. Relevance of the present research work	15
	15
Chapter 2 Survey of MT systems in India and Abroad	15
2.0. Introduction	15
2.1 Machine Translation	16
2.1.1 Machine Translation System for non Indian languages	17
2.1.2 Machine Translation Systems for Indian languages	28
2.2 History of Machine Translation	37
2.3 Need for MT	42
2.4 Problems in MT	42
2.5 Types of Machine Translation Systems	43
2.6 Different Approaches used for Machine Translation	44
2.6.1 Linguistic or Rule-Based Approaches:	44
2.6.1.1 Direct MT System:	45
2.6.1.2 Interlingua Machine Translation	47
2.6.1.3 Transfer based MT	49
2.6.2 Non-Linguistic Approaches	50
2.6.2.1 Dictionary Based Approach	50
2.6.2.2 Empirical or Corpus Based Approaches	51
2.6.2.2.1 Example Based Approach	51
2.6.2.2.2 Statistical Approach	52
2.6.3 Hybrid Machine Translation Approach	54
2.7 Categories of Machine Translation System	54
2.7.1 Fully Automated Machine Translation System	55
2.7.2 Machine Aided Translation System	55
2.7.3 Terminology Data Banks	56
2.8 Advantages of Statistical Machine Translation over Rule Based Machine Translation	56
2.9 Applications of Machine Translation	58
2.10 Summary	63
Chapter 3 Creation of Parallel Corpus	63
3.0 Introduction	63
3.1 Pre-Electronic corpus	63
3.2. Corpus in the present day context	64
3.2.1. Sampling and representativeness	65
3.2.2. Finite size	66
3.2.3. Machine-redable form	67
3.2.4. A standard reference	67

3.3. Classification of the corpus	68
3.3.1. Genere of text	68
3.3.2. Nature of data	68
3.3.3. Type of text	69
3.3.4. Purpose of design	70
3.3.5. Nature of application	70
3.3.5.1. Alined corpus	70
3.3.5.2. Parallel copus	71
3.3.5.3. Reference corpus	71
3.3.5.4. Comparable corpus	71
3.3.5.5. Opportunistic corpus	71
3.4. Generation of written corpus	72
3.4.1. Size of corpus	72
3.4.2. Represetativeness of texts	73
3.4.3. Nativity	73
3.4.4. Determination of target users	75
3.4.5. Selection of time-span	76
3.4.6. Selection of texts type	76
3.4.7. Method of data sampling	77
3.4.8. Method of data input	77
3.4.9. Hardware requirement	78
3.4.10. Management of corpus files	79
3.4.11. Method of corpus sanitation	79
3.4.12. Problem of copy right	80
3.5. Corpus processing	80
3.5.1. Frequency study	81
3.5.2. Word sorting	81
3.5.3. Concordance	82
3.5.4 Lexical Collocation	82
3.5.5 Key Word In Context (Kwic)	82
3.5.6 Local Word Grouping (Lwg)	83
3.5.7 Word Processing	83
3.5.8 Tagging	85
3.6. Parallel corpora	85
3.6.1. Parallel corpora types	86
3.6.2. Examples of parallel corpora	88
3.6.3. Applications of parallel corpora	89
3.6.4. Corpora creation in Indian languages	91
3.6.4.1. POS tagged corpora	92
3.6.4.2. Chunked corpora	92
3.6.4.3. Semantically tagged corpora	92
3.6.4.4. Syntacitc tree bank	93
3.6.4.5. Souces for parallel corpora	94
3.6.4.6. Tools	94
3.6.5. Creating multilingual parallel corporat for Indian languages	95
3.6.5.1. Creating the source text	97
3.6.5.2. Domain of corpus	97
3.6.5.2.1. Health Domain	97

3.6.5.2.2. Tourism domain	98
3.6.5.3. Data storage, maintenance and dissemination	98
3.6.5.4. Parallel corpus creation	99
3.6.5.5. POS Annotation	99
3.6.5.5.1 POS Tagset	99
3.6.5.5.1.1 Principles for Designing Linguistic Standards for Corpora Annotation	100
3.6.5.5.2 Super Set of POS Tags	101
3.6.5.5.3 Super Set of POS Tags for Indian Languages	101
3.6.5.5.4 Manual POS Annotation	101
3.6.6. Creation of parallel corpus for the SMT system	102
3.6.6.1. Corpus collection	103
3.6.6.2. Compilation of parallel corpora	104
3.6.6.3. Alinment of the parallel corpus	104
3.6.6.4. Sentence alignment	105
3.6.6.5. Word alignment	106
3.7. Summary	108
Chapter 4 Parallel Structure of English and Tamil Language	108
4.0 Introduction	108
4.1. Parallel sentential structures in English and Tamil	109
4.1.1. Prallels affirmative sentences	116
4.1.2. Parallels in interrogative sentences	117
4.1.2.1. Parallels in yes-no questions	118
4.1.2.2. Parallels of wh-questions	120
4.1.3. Parallels in negative sentences	122
4.1.3.1. Parallels in negation in equative sentences	122
4.1.3.2. Parallels in negation in non-equative sentences	122
4.1.3.3. Paralles in negative pronouns and determiners	123
4.1.4. Parallesl in imperative sentence	125
4.2. Parallel clause structures of English and Tamil	126
4.2.1. Parallels in nomainal/complement clause	130
4.2.2. Paralles in Adverbial clauses	131
4.2.3. Parallels in Adjectival clauses	134
4.2.4. Parallels in comparative clauses	136
4.2.4.1. Parallels in comparative clause of quality	137
4.2.4.2. Parallels in comparative clause of quantity	137
4.2.4.3. Parallels in comparative clause of adverbs	138
4.2.5. Parallels in co-ordination	138
4.3. Parallel structures of English and Tamil pharases	139
4.3.1. Paralles in noun phrases	139
4.3.1.1. Parallels in demonstratives	139
4.3.1.2.Parallels in quantifiers	141
4.3.1.3. Parallels in genitive pharase	142
4.3.2. Parallel structures in verb phrase	143
4.3.2.1. Parallels in complex verbal forms denoting tense, mood and aspect	144
4.3.2.2. Parallels in verb patterns	151
4.3.3. Parallels in adjectival pharases	158

4.3.4. Paralles in adverbial phrase	159
4.3.5. Parallels in adpositional phrases	165
4.3.6. Parallels in phrasal co-ordination	170
4.4. Summary	171
Chapter 5 English to Tamil Machine Translation System by using Parallel corpus	172
5.0 Introduction	172
5.1 On the subject of SMT	172
5.1.1. Statistical Machine Translation and the Noisy Channel Model	173
5.1.2 Advantages of SMT	174
5.1.3 Challenges with statistical machine translation	175
5.2 The Components of Statistical Machine Translation	176
5.2.1 Language Model	177
5.2.2 Translation Model	178
5.2.2.1 Expectation Maximization	178
5.2.2.2 Different Translation Models	179
5.2.2.2.1 Word-based Translation Model	179
5.2.2.2.2 Phrase-based Translation Model	180
5.2.2.2.3. Factored Translation Model	182
5.2.3 The Statistical Machine Translation Decoder	183
5.3 Tools used for implementation of SMT System	184
5.3.1 Language Model (LM) tools	185
5.3.2 Translation Model Tools	186
5.3.3 Decoder Tools	186
5.4 Existing Statistical MT Systems	187
5.5 Problem Statement	188
5.5.1 Gap Analysis	188
5.6 Development of Corpus	189
5.7 Architecture of English to Tamil Statistical Machine Translation System	189
5.7.1 Architecture for Language Model	190
5.7.2 Architecture for Translation Model	190
5.7.3 Architecture for Decoder	191
5.8 Preparation of Data	191
5.8.1 Tokenizing the corpus	191
5.8.2 Filtering out long sentences	192
5.8.3 Lowercasing data	193
5.9 Generating Language Model	194
5.9.1 Installation of SRILM	196
5.10 Generating Translation Model	198
5.10.1 Installation of *GIZA++*	198
5.11 Generating Decoder	199
5.11.1 Installation of Moses	199
5.11.2 Training Moses decoder	201
5.11.3 Tuning Moses decoder	202
5.11.4 Running Moses decoder	203
5.12 EXPERIMENTAL FRAMEWORK	210
5.12.1 English – Tamil Phrase Based Statistical Machine Translation	210

System	
5.12.2 Proposed System Architecture	210
5.13 Implementation	212
5.13.1 Development of Bilingual Corpus for English –Tamil language pair	212
5.13.2 Development of Monolingual Corpus for Tamil language	212
5.13.3 Pre-processing of Corpus	213
5.13.4 Building Language Model	213
5.13.5 Building Phrase-Based Translation Model	214
5.13.6 Tuning	216
5.13.7 Experimental Results	217
5.14 Handling Idioms and Phrasal Verbs in Machine Translation	217
5.14.1 Overview	217
5.14.2 Idioms and Phrasal Verbs in Machine Translation	217
5.14.3 Phrasal Verbs and Idioms – An Overview	217
5.14.4 Challenges in Handling Idioms and Phrasal Verbs	221
5.14.5 Implementation	222
5.14.5.1 Rule Based Machine Translation System	224
5.14.5.2 Factored Statistical Machine Translation System	225
5.14.6 Experimental Results	226
5.15 Automated Factored Information Generation for English and Tamil	227
5.15.1 Factor Annotator for English	228
5.15.2 Factor Annotator for Tamil	229
5.16 Beyond Standard Statistical Machine Translation	230
5.16.1 Factored Translation Models	230
5.16.2 Syntax based Translation Models	232
5.16.3 On-going Research	233
5.17. Summary	233
Chapter 6: Conclusion	233
Appendix 1 A sample of English and Tamil parallel corpus	238
BIBLIOGRAPHY	249

CHAPTER 1
INTRODUCTION

1.1 Motivation

Machine translation is the task of translating the text in source language to target language, automatically. Machine translation can be considered as an area of applied research that draws ideas and techniques from linguistics, computer science, artificial intelligence, translation theory, and statistics. Even though machine translation was envisioned as a computer application in the 1950‘s and research has been made for 60 years, machine translation is still considered to be an open problem.

The demand for machine translation is growing rapidly. As multilingualism is considered to be a part of democracy, the European Union funds EuroMatrixPlus, a project to build machine translation system for all European language pairs, to automatically translate the documents to 23 official languages, which were being translated manually. Also as the United Nations is translating a large number of documents into several languages, the UN has created bilingual corpora for some language pairs like Chinese – English, Arabic–English which are among the largest bilingual corpora distributed through the Linguistic Data Consortium. In the World Wide Web, as around 20% of web pages and other resources are available in their national languages, machine translation can be used to translate these web pages and resources to the required language in order to understand the content in those pages and resources, thereby decreasing the effect of language as a barrier of communication.

In a linguistically diverged country like India, machine translation is an important and most appropriate technology for localization. Human translation in India can be found since the ancient times which are being evident from the various works of philosophy, arts, mythology, religion and science which have been translated among ancient and modern Indian languages. Also, numerous classic works of art, ancient, medieval and modern, have also been translated between

European and Indian languages since the 18th century. As of now, human translation in India finds application mainly in the administration, media and education, and to a lesser extent, in business, arts and science and technology.

India has 22 constitutional languages, which were written in 10 different scripts. Hindi is the official language of the India. English is the language which is most widely used in the media, commerce, science and technology and education. Many of the states have their own regional language, which is either Hindi or one of the other constitutional languages. Only about 5% of the population speaks English.

In such a situation, there is a big market for translation between English and the various Indian languages. Currently, the translation is done manually. Use of automation is largely restricted to word processing. Two specific examples of high volume manual translation are -translation of news from English into local languages, translation of annual reports of government departments and public sector units among, English, Hindi and the local language. Many resources such as news, weather reports, books, etc., in English are being manually translated to Indian languages. Of these, News and weather reports from all around the world are translated from English to Indian languages by human translators more often. Human translation is slow and also consumes more time and cost compared to machine translation. It is clear from this that there is large market available for machine translation rather than human translation from English into Indian languages. The reason for choosing automatic machine translation rather than human translation is that machine translation is better, faster and cheaper than human translation.

Tamil, a Dravidian language spoken by around 72 million people is the official language of Tamil Nadu state government of India. Tamil in its eagerness to gather information from English resort to build English-Tamil machine translation systems. Many English-Tamil machine translation systems are getting built, but none could serve the ambitious need of Tamil. This work is intended pursue this work in a new perspective.

1.2 Issues in the research

Natural language processing has many challenges, of which the biggest is the inherent ambiguity of natural language. Machine translation systems have to deal with ambiguity, and various other natural language phenomena. In addition, the linguistic diversity between the source and target language makes machine translation a bigger challenge. This is particularly true for widely divergent languages such as English and Tamil. The major structural difference between English and Tamil can be summarized as follows. English is a highly positional language with rudimentary morphology, and default sentence structure as SVO. Tamil is highly inflectional, with a rich morphology, relatively free word-order, and default sentence structure as SOV. In addition, there are many stylistic differences. For example, it is common to see very long sentences in English, using abstract concepts as the subjects of sentences, and stringing several clauses together. Such constructions are not natural in Tamil, and this leads to major difficulties in producing good translations. Compared to English, Tamil is rich in morphology and is an agglutinative language. As it is recognized all over the world, with the current state of art in machine translation, it is not possible to have fully automatic, high quality, and general-purpose machine translation. Practical systems need to handle ambiguity and the other complexities of natural language processing, by relaxing one or more of the above dimensions.

The present thesis work addresses the above problem with the new perspective of building a statistical machine translation system for English to Tamil using parallel corpus. The accuracy of the translation in the statistical approach mainly depends on the size of the bilingual corpus of English- Tamil language pair and also on the size of the monolingual corpus of the target language. Handling the phrasal verbs and idioms is one of the major issues in English-Tamil machine translation system. Also determining the morph lexical information from the bilingual and monolingual corpus in order to generate a factored bilingual and monolingual corpus, which have been done manually has to be automated so as to reduce the time and cost involved in generating the factored corpus from the normal bilingual and monolingual corpus. The above problems can be addressed by determining a way to automate the generation factored information for both the source and target language, determine a technique to handle the phrasal verbs and idioms and

increasing the size of the bilingual corpus of English-Tamil language pair and the size of monolingual corpus of the target language, Tamil.

Most of the content available in digital format is in English language. The content shown in English must be presented in a language which can be understood by the intended audience. There is large section of population at both national and state level who cannot comprehend English language. It has brought about language barrier in the side lines of digital age. Machine Translation (MT), can overcome this barrier. In this thesis, a proposed Statistical Based Machine Translation system for translating English text to Tamil language has been proposed. English is the source language and the Tamil is the target language.

1.3 Aims and Objectives of the work

Here in this research work it is proposed to design a machine translation system based on the statistical approach, along with a pre-processing technique to handle phrasal verbs and idioms in both factored statistical and rule based machine translation system and a module to generate factored information for the factored statistical machine translation system for English-Tamil from the raw bilingual corpus of English-Tamil language pair. The main objectives of the thesis work are

- To understand the Language Model (LM), Translation Model (TM), and Decoding stages of SMT.
- To create a LM for Tamil with use of SRI's LM language model.
- To create a TM model with use of GIZA++ software.
- To generate Tamil sentences with use of Moses software.
- To evaluate and test the system.
- To increase the size of the bilingual corpus of English – Tamil language pair and the size of the monolingual corpus of target language, Tamil.
- To develop a module that generates the factored information for the source language, English and the target language, Tamil for training the Factored Statistical machine translation system for English-Tamil.

- To develop a pre-processing technique to handle the phrasal verbs and idioms.

1.4. Methodology

The present research work makes use of the statistical machine translation approach for English to Tamil rather than the other approaches of machine translation such as rule based and example based. Because of the complexities in other approaches which will be discussed briefly in the later chapters.

English to Tamil language translation is built here by making use of Statistical Machine Translation (SMT). Main goal of this system is to undertake translation with minimum human efforts. There are many tools pertaining to LM, TM, decoder for undertaking SMT. SMT has three major parts of the system, Language Model, Translation Model and searching (decoder). The LM computes the probabilities with respect to the target language. The TM computes the probabilities regarding the substitution of target language word with source language word. For development of LM, SRI international's SRILM Language Model toolkit is used. GIZA++ is used for creation of Translation Model. For decoding stage, Moses software has been used. The system is based upon Linux operating system. It will accept English sentence from the terminal and produce output in Tamil.

1.5 Previous research works

There are many attempts in translating English into Tamil using machine. The department of Information technology, Govt. of India has started a project called Technology Development for Indian Languages (TDIL) in 1991 and supporting a number of research institutes in the country for the development of all the 22 scheduled languages.

- Anusaraka Project: An MAT project was started at IIT, Kanpur for translation among Indian languages based on Paninian grammatical formalism. The transfer at the word level exploits the similarities found in the structure of Indian languages.

- Angla Bharati Project: An MAT system to transfer English into Hindi was launched at IIT, Kanpur.
- MAT of Standard Documents: It is a domain specific translation system, which aims to transfer English text into Hindi. It basically follows Angla Bharati approach.
- MAT from English to Hindi: It is an ongoing project at CDAC Pune. It concentrates on the translation of administrative languages.
- Software to translate texts from English to Tamil: A project headed by Mr. Duraipandi.
- Siva & Shakti MT aids prepared by IISC, Bangalore and IIIT, Hyderabad.
- DIT is supporting English to Indian language machine translation project. Under project English language to Indian Language (ELMT) project a system called ANUVADAKSH is getting built. The first phase is over and the second phase is going on. Under the scheme Amrita University, Coimbatore is building English-Tamil machine translation system.

Tamil university has built a translation system to translate between Russian language and Tamil.

A PhD thesis submitted by Kamakshi which has been published in the title "Preliminaries to the preparation of a Machine Translation Aid to Translate Linguistics Texts written in English to Tamil." (Kamashi and Rajendran 2004) is an extensive work based on transfer approach. Kamshi discusses elaborately about the structural differences of English and Tamil and she has made use of lexical-transfer approach to build an aid to translate English text books in English into Tamil. She has listed a series of transfer rules and build a elaborate bilingual dictionary to serve her purpose.

The details of the previous works are given elaborately in the second chapter.

1.6 Chapterization

- Chapter 1 briefly introduces the topic of the research work. It discusses about the aims and objectives, methodology, earlier works in the field of investigation and the uses of the present research work.

- Chapter 2 presents a literature survey of the machine translation systems and the theoretical background of machine translation and its various approaches.

- Chapter 3 presents the details of creation of parallel corpus for English-Tamil SMT system.

- Chapter 4 presents an overview on parallel structures of English and Tamil language.

- Chapter 5 presents an experimental framework on implementation and results of a phrase-based statistical machine translation for English-Tamil, a technique to handle phrasal verbs and idioms in machine translation and design of automated wrappers for English and Tamil to annotate English and Tamil sentences with factors such as lemma, part of speech information and morphology.

- Chapter 6 presents the conclusion and possible future work addressed by this thesis.

1.7 Relevance of the present research work

Machine translation is the order of the day. Building rule based machine translation systems are time consuming and uneconomical. So the best alternative is to build Statistical based machine translation system using parallel corpus. The present work is only a starting point. With the availability of huge English-Tamil parallel corpus the system will improve and supersede Google English-Tamil on-line translation system which is founded on the same ground.

Chapter -2

Survey of MT systems in India and abroad

2.0 Introduction

The technology is reaching new heights, right from conception of ideas up to the practical implementation. It is important, that equal emphasis is put to remove the

language divide which causes communication gap among different sections of societies. Natural Language Processing (NLP) is the field that strives to fill this gap. Machine Translation (MT) mainly deals with transformation of one language to another. Coming to the MT scenarios in India, it has enormous scope due to many regional languages of India. It is pertinent that majority of the population in India are fluent in regional languages such as Hindi, Punjabi *etc.*. Given such a scenario, MT can be used to provide an interface of regional language. This chapter aims to survey MT systems in India and aboard along with a brief history of MT.

2.1 Machine Translation

Machine translation is one of the major, oldest and the most active area in natural language processing. The word 'translation' refers to transformation of one language into other. Machine Translation is the process of using computers to automate some or all of the process of translation from one language to another. It is an area of applied research that draws ideas and techniques from linguistics, computer science, artificial intelligence, translation theory, and statistics. It is a focused field of research in linguistic concepts of syntax, semantics, pragmatics and discourse, computational-linguistic approaches such as parsing algorithms, semantic and pragmatic clarification and text generation, descriptive linguistics that deals with lexicon and language rules for particular languages and modeling human knowledge representation and manipulation. Research began in this field as early as in the late 1940s, and numerous methods some based on extensive linguistic theories and some ad-hoc have been tried over the past five decades.

Machine translation can also be defined as, the application of computers to the task of translating texts from one natural language to another. Today a number of systems are available that are capable of producing translations which, even though not perfect, is of sufficient quality to use in a number of specific domains. In the process of translation, which either carried out manually or automated through machines, the context of the text in the source language when translated must convey the exact context in the target language. While seeing from the surface, this seems straightforward, but it is far more difficult. Translation is not a just a word level replacement. A translator, either a machine or human, must interpret and analyse all the

elements in the text. Also he should be familiar with all the issues during the translation process and must know how to handle it. This requires widespread knowledge in grammar, sentence structure, meanings, etc., in the source and target languages, also with understanding with each language‘s culture in order to handle idioms and phrases which gets originated from different culture and becomes an important issue that affect the accuracy of the translation.

It will be a great challenge for human to face various challenges in the designing a machine translation system, proficient of translating sentences by taking consideration of all the required information to perform translation. Even though, no two individual human translators can generate similar translations of the same text in the same language pair and it may take several revisions to make the translation perfect. Hence it will be a greater challenge for humans to design a fully automated machine translation system to produce quality translations.

2.1.1 Machine Translation System for non Indian languages

Various Machine Translation systems have already been developed for most of the commonly used natural languages. This section briefly discusses some of the existing Machine Translation systems and the approaches that have been followed (Hutchins, 1986, 1994, 2005; Solcum 1985).

Georgetown Automatic Translation (GAT) System (1952), developed by Georgetown University, used direct approach for translating Russian texts (mainly from physics and organic chemistry) to English. The GAT strategy was simple word- for-word replacement, followed by a limited amount of transposition of words to result in something vaguely resembling English. There was no true linguistic theory underlying the GAT design. It had only six grammar rules and 250 items in its vocabulary. The translation was done using IBM 701 mainframe computer. Georgetown University and IBM jointly conducted the Georgetown-IBM experiment in 1954 for more than sixty Russian sentences into English. The experiment was a great success and ushered in an era of Machine Translation research. The Georgetown MT project was terminated in the mid-60s.

CETA (1961), included the linguistic theory unlike GAT, for translating Russian into French. It was developed at Grenoble University in France. It is based on Interlingua approach with dependency-structure analysis of each sentence at the grammatical level and transfer mapping from one language-specific meaning representation at the lexical level. During the period of 1967-71, this system was used to translate about 4,00,000 words of Russian mathematics and physics texts into French. It was found that it fails for those sentences for which complete analysis cannot be derived. In 1971, new and improved system GETA based on the limitations of CETA was developed.

METAL (Mechanical Translation and Analysis of Languages) (1961) was developed at Linguistics Research Center, University of Texas for German into English. The system used indirect Machine Translation approach using Chomsky‘s transformational paradigm. Indirect translation was performed in 14 steps of global analysis, transfer, and synthesis. The performance and accuracy of the system was moderate.

The Mark II (1964), a direct translation approach based Russian to English MT System for U.S. Air Force. It was developed by IBM Research Center. Translation was word by word, with occasional backtracking, Each Russian item (either stem or ending) in the lexicon was accompanied by its English equivalent and grammatical codes indicating the classes of stems and affixes that could occur before and after it. In addition to lexical entries, processing instructions were also intermixed in the dictionary: “control entries” relating to grammatical processes (forward and backward skips), and also instructions relating to loading and printing routines. There were some 25,000 such “control entries” included in the dictionary. This contained 150,000 entries at the World‘s Fair demonstration, and 180,000 in the USAF version. A third of the entries were phrases, and there was also an extensive system of micro glossaries. An average translation speed of 20 words per second was claimed. The examples of Russian-English translations at the World‘s Fair were reasonably impressive (Bowers & Fisk (1965)). The Russian-English translations produced by Mark II were often rather crude and sometimes far

from satisfactory. The limitations of word by word translation are more evident in the evaluation reports submitted by Pfafflin (1965), Orr & Small (1967), ALPAC (1966). An evaluation, MT research at the IBM Research Center ceased in 1966 (Roberts & Zarechnak 1974). As one of the first operational MT systems, the IBM Russian-English system has a firm place in the history of MT. It was installed in the USAF‘s Foreign Technology Division at the Wright-Patterson Air Force Base, Dayton, Ohio, where it remained in daily operation until 1970.

LOGOS (1964), a direct Machine Translation system for English-Vietnamese language pair was initially developed by US Private firm Logos Corporation. Logos analyzes whole source sentences, considering morphology, meaning, and grammatical structure and function. The analysis determines the semantic relationships between words as well as the syntactic structure of the sentence. Parsing is only source language-specific and generation is target language-specific. Unlike other commercial systems the Logos system relies heavily on semantic analysis. This comprehensive analysis permits the Logos system to construct a complete and idiomatically correct translation in the target language. This Internet-based system allows 251 users to submit formatted documents for translation to their server and retrieve translated documents without loss of formatting. In 1971, It was used by the U.S. Air Force to translate English maintenance manuals for military equipment into Vietnamese. Eventually, LOGOS forged an agreement with the Wang computer company that allowed the implementation of the German-English system on Wang office computers. This system reached the commercial market, and has been purchased by several multi-national organizations (e.g., Nixdorf, Triumph- Adler, Hewlett-Packard). The System is also available for English-French, English-German language pairs.

TAUM-AVIATION (1965), a transfer approach based English - French MT System for weather forecasts. It was developed at University of Montreal. After short span of time, the domain for translation shifted to translating aviation manuals by adding semantic analysis module to the system. The TAUM-AVIATION system is based on a typical second generation design

(Isabelle et al. 1978, Bourbeau 1981). The translation is produced indirectly, by means of an analysis/transfer/synthesis scheme. The overall design of the system is based on the assumption that translation rules should not be applied directly to the input string, but rather to a formal object that represents a structural description of the content of this input. Thus, the source language (SL) text (or successive fragments of it) is mapped onto the representations of an intermediate language, (also called normalized structure) prior to the application of any target language-dependent rule. In this system, the dictionaries list only the base form of the words (roughly speaking, the entry form in a conventional dictionary). In March 1981, the source language (English) dictionary included 4054 entries; these entries represented the core vocabulary of maintenance manuals, plus a portion of the specialized vocabulary of hydraulics. Of these, 3280 had a corresponding entry in the bilingual English-French dictionary. The system was evaluated and the low accuracy of the translation by the system forced the Canadian Government to cancel the funding and thus TAUM project in 1981.

SYSTRAN (1968) is a direct Machine Translation system developed by Huchins and Somers. The system was originally built for English-Russian Language Pair. In 1970, SYSTRAN System installation at United States Air Force (USAF) Foreign Technology Division (FTD) at Wright-Patterson Air Force Base, Ohio, replaced IBM MARK-II MT System and is still operational. Large number of Russian scientific and technical documents were translated using this system. The quality of the translations, although only approximate, was usually adequate for understanding content. In 1974, NASA also selected SYSTRAN to translate materials relating to the Apollo-Soyuz collaboration, and in 1976, EURATOM replaced GAT with SYSTRAN. The Commission of the European Communities (CEC) purchased an English-French version of SYSTRAN for evaluation and potential use. Unlike the FTD, NASA, and EURATOM installations, where the goal was information acquisition, the intended use by CEC was for information dissemination - meaning that the output was to be carefully edited before human consumption. The quality for this purpose was not adequate but improved after adding lexicon entries specific to CEC related translation tasks. Also in 1976, General Motors of

Canada acquired SYSTRAN for translation of various manuals (for vehicle service, diesel locomotives, and highway transit coaches) from English into French on an IBM mainframe. GM's English-French dictionary had been expanded to over 1,30,000 terms by 1981 (Sereda 1982). GM purchased an English-Spanish version of SYSTRAN, and began to build the necessary [very large] dictionary. Sereda (1982) reported a speed-up of 3-4 times in the productivity of his human translators. Currently, SYSTRAN System is available for translating in 29 language pairs.

CULT(Chinese University Language Translator)(1968), is an interactive online MT System based on direct translation strategy for translating Chinese mathematics and physics journals into English. Sentences are analyzed and translated one at a time in a series of passes. After each pass, a portion of the sentence is translated into English. The CULT includes modules like source text preparation, input via Chinese keyboard, lexical analysis, syntactic and semantic analysis, relative order analysis, target equivalence analysis, output and output refinement. CULT is a successful system but it appears somewhat crude in comparison to interactive systems like ALPS and Weidner.

ALPS (1971), a direct approach based English into French, German, Portuguese and Spanish for Mormon ecclesiastical texts. It was developed at Brigham Young University. It was started with an aim to develop fully automatic MT System but later in 1973, it became Machine Aided System. It is an Interactive Translation System that performs global analysis of sentences with human assistance, and then performs indirect transfer again with human assistance. But this project was not successful and hence not operational.

The METEO (1977) , is the world‘s only example of a truly fully automatic MT System for Canadian Meterological Centre‘s(CMC‘s) nationwide weather communication networks. METEO scans the network traffic for English weather reports, translates them directly into French, and sends the translations back out over the communications network automatically. This system is based on the TAUM technology as discussed earlier. It is probably

the first MT system where translators had involved in all phases of the design, development and refinement. Rather than relying on post-editors to discover and correct errors, METEO detects its own errors and passes the offending input to human editors and output deemed correct by METEO is dispatched without human intervention. This system correctly translates 90-95%, shuttling the other 5-10% to the human CMC translators.

An English Japanese Machine Translation System (1982) developed by Makoto Nagao et. al. The title sentences of scientific and engineering papers are analyzed by simple parsing strategies. Title sentences of physics and mathematics of some databases in English are translated into Japanese with their keywords, author names, journal names and so on by using fundamental structures. The translation accuracy for the specific areas of physics and mathematics from INSPEC database was about 93%.

RUSLAN (1985), a direct Machine Translation system between closely related languages Czech and Russian, by Hajic J, for thematic domain, the domain of operating systems of mainframes. The system used transfer based architecture. This project started in 1985 at Charles University, Prague in cooperation with Research Institute of Mathematical Machines in Prague. It was terminated in 1990 due to lack of funds. The system was rule based, implemented in Colmerauer‘s Q-Systems. The system had a main dictionary of about 8,000 words, accompanied by transducing dictionary covering another 2,000 words. The typical steps followed in the system are Czech morphological analysis, syntactico semantic analysis with respect to Russian sentence structure and morphological synthesis of Russian. Due to close language pair, a transfer-like translation scheme was adopted with many simplifications. Also many ambiguities are left unresolved due to the close relationship between Czech and Russian. No deep analysis of input sentences was performed. The evaluations of results of RUSLAN showed that roughly 40% of the input sentences were translated correctly, about 40% of input sentences with minor errors correctable by human post-editor and about 20% of the input required substantial editing or re-translation. There are two main factors that caused a deterioration of the translation. The first factor was

the incompleteness of main dictionary of the system and second factor was the module of syntactic analysis of Czech. RUSLAN is a unidirectional system dealing with one pair of language Czech to Russian.

PONS (1995) , an experimental interlingua system for automatic translation of unrestricted text, constructed by Helge Dyvik, Department of Linguistics and Phonetics, University of Bergen. 'PONS' is in Norwegian an acronym for "Partiell Oversettelse mellom Nærstående Språk" (Partial Translation between Closely Related Languages). PONS exploits the structural similarity between source and target language to make the shortcuts during the translation process. The system makes use of a lexicon and a set of syntactic rules. There is no morphological analysis. The lexicon consists of a list of entries for all word forms and a list of stem entries, or ‗lexemes'. The source text is divided into substrings at certain punctuation marks, and the strings are parsed by a bottom-up, unification-based active chart parser. The system had been tested on translation of sentence sets and simple texts between the closely related languages Norwegian and Swedish, and between the more distantly related English and Norwegian.

interNOSTRUM (1999) is a bidirectional Spanish-Catalan Machine Translation system. It was developed by Marote R.C. et al. It is a classical indirect Machine Translation system using an advanced morphological transfer strategy. Currently it translates ANSI, RTF (Microsoft's Rich Text Format) and HTML texts. The system has eight modules: a deformatting module which separates formatting information from text, two analysis modules (morphological analyzer and part-of-speech tagger), two transfer modules (bilingual dictionary module and pattern processing module) and two generation modules (morphological generator and post-generator), and the reformatting module which integrates the original formatting information with the text. This system achieved great speed through the use of finite-state technologies. Error rates range around 5% in Spanish-Catalan direction when newspaper text is translated and are somewhat worse in the Catalan-Spanish direction. The Catalan to Spanish is less satisfactory as to vocabulary coverage and accuracy.

ISAWIKA!(1999) is a transfer-based English-to-Tagalog MT system that uses ATN (Augmented Transition Network) as the grammar formalism. It translates simple English sentences into equivalent Filipino sentences at the syntactic level.

English-to-Filipino MT system (2000) is a transfer based MT System that is designed and implemented using the lexical functional grammar (LFG) as its formalism. It involves morphological and syntactical analyses, transfer and generation stages. The whole translation process involves only one sentence at a time.

Tagalog-to-Cebuano Machine Translation System (T2CMT)(2000) is a uni-directional Machine Translation system from Tagalog to Cebuano. It has three stages: Analysis, Transfer and Generation. Each stage uses bilingual from Tagalog to Cebuano lexicon and a set of rules. The morphological analysis is based on TagSA (Tagalog Stemming Algorithm) and affix correspondence-based POS (part-of-speech) tagger. The author describes that a new method is used in the POS-tagging process but does not handle ambiguity resolution and is only limited to a one-to-one mapping of words and parts-of-speech. The syntax analyzer accepts data passed by the POS tagger according to the formal grammar defined by the system. Transfer is implemented through affix and root transfers. The rules used in morphological synthesis are reverse of the rules used in morphological analysis. T2CMT has been evaluated, with the Book of Genesis as input, using GTM (General Text Matcher), which is based on Precision and Recall. Result of the evaluation gives a score of good performance 0.8027 or 80.27% precision and 0.7992 or 79.92% recall.

Turkish to English Machine Translation system(2000) is a hybrid Machine Translation system by combining two different approaches to MT. The hybrid approach transfers a Turkish sentence to all of its possible English translations, using a set of manually written transfer rules. Then, it uses a probabilistic language model to pick the most probable translation out of this set. The system is evaluated on a test set of Turkish sentences, and

compared the results to reference translations. The accuracy comes out to be about 75.6%.

CESILKO(2000), is a Machine Translation system for closely related Slavic language pairs, developed by HAJIC J, HRIC J K. and UBON V. It has been fully implemented for Czech to Slovak, the pair of two most closely related Slavic languages. The main aim of the system is localization of the texts and programs from one source language into a group of mutually related target languages. In this system, no deep analysis had been performed and word-for-word translation using stochastic disambiguation of Czech word forms has been performed. The input text is passed through different modules namely morphological analyzer, morphological disambiguation, Domain related bilingual glossaries, general bilingual dictionary, and morphological synthesis of Slovak. The dictionary covers over 7, 00,000 items and it is able to recognize more than 15 million word-forms. The system is claimed to achieve about 90% match with the results of human translation, based on relatively large test sample. Work is in progress on translation for Czech-to-Polish language pairs.

Bulgarian-to-Polish Machine Translation system (2000), has been developed by S. Marinov. This system has been developed based on the approach followed by PONS discussed above. The system needs a grammar comparison before the actual translation begins so that the necessary pointers between similar rules are created and system is able to determine where it can take a shortcut. The system has three modes, where mode 1 and 2 enable system to use the source language constructions and without making a deeper semantic analysis to translate to the target language construction. Mode 3 is the escape hatch, when the Polish sentences have to be generated from the semantic representation of the Bulgarian sentence. The accuracy of the system has been reported to be 81.4%.

Tatar (2001), a Machine Translation system between Turkish and Crimean, developed by Altintas K. et al., used finite state techniques for the translation process. It is in general disambiguated word for word translation. The system

takes a Turkish sentence, analyses all the words morphologically, translates the grammatical and context dependent structures, translates the root words and finally morphologically generates the Crimean Tatar text. One-to-one translation of words is done using a bilingual dictionary between Turkish and Crimean Tatar. The system accuracy can be improved by making word sense disambiguation module more robust.

Antonio M. Corbí-Bellot et. al. (2005) developed the open source shallow-transfer Machine Translation (MT) engine for the Romance languages of Spain (the main ones being Spanish, Catalan and Galician). The Machine Translation architecture uses finite-state transducers for lexical processing, hidden Markov models for part-of-speech tagging, and finite-state based chunking for structural transfer. The author claims that, for related languages such as Spanish, Catalan or Galician, a rudimentary word-for-word MT model may give an adequate translation for 75% of the text, the addition of homograph disambiguation, management of contiguous multi-word units, and local reordering and agreement rules may raise the fraction of adequately translated text above 90%.

Carme Armentano-oller et. al (2005) extended the idea of A.M.Corbi-Bellot et. al. and developed an open source Machine Translation tool box which includes (a) the open-source engine itself, a modular shallow transfer Machine Translation engine suitable for related languages (b) extensive documentation specifying the XML format of all linguistic (dictionaries, rules) and document format management files, (c) compilers converting these data into the high speed format used by the engine, and (d) pilot linguistic data for Spanish-Catalan and Spanish-Galician and format management specifications for the HTML, RTF and plain text formats. They use the XML format for linguistic data used by the system. They define five main types of formats for linguistic data i.e. dictionaries, tagger definition file, training corpora, structural transfer rule files and format management files.

Apertium (2005), developed by Carme Armentano-oller et. al is an open-source shallow-transfer Machine Translation (MT) system for the [European]

Portuguese ↔ Spanish language pair. This platform was developed with funding from the Spanish government and the government of Catalonia at the University of Alicante. It is a free software and released under the terms of the GNU General Public License. Apertium originated as one of the Machine Translation engines in the project OpenTrad and was originally designed to translate between closely related languages, although it has recently been expanded to treat more divergent language pairs (such as English–Catalan). Apertium uses finite-state transducers for all lexical processing operations (morphological analysis and generation, lexical transfer), hidden Markov models for part-of-speech tagging, and multi-stage finite-state based chunking for structural transfer. For Portuguese–Spanish language pair, promising results are obtained with the pilot open-source linguistic data released which may easily improve (down to error rates around 5%, and even lower for specialized texts), mainly through lexical contributions from the linguistic communities involved.

ga2gd (2006), a robust Machine Translation system, developed by Scannell K.P., between Irish and Scottish Gaelic despite the lack of full parsing technology or pre-existing bilingual lexical resources. It includes the modules Irish standardization, POS Tagging, stemming, chunking, WSD, Syntactic transfer, lexical transfer, and Scottish post processing. The accuracy has been reported to be 92.72%.

SisHiTra(2006) is a hybrid Machine Translation system from Spanish to Catalan. It was developed by Gonzalez et. al. This project tried to combine knowledge-based and corpus-based techniques to produce a Spanish-to-Catalan Machine Translation system with no semantic constraints. Spanish and Catalan are languages belonging to the Romance language family and have a lot of characteristics in common. SisHiTra makes use of their similarities to simplify the translation process. A SisHiTra future perspective is the extension to other language pairs (Portuguese, French, Italian, etc.). The system is based on finite state machines. It has following modules: preprocessing modules, generation module, disambiguation module and post-

processing module. The word error rate is claimed to be 12.5% for SisHiTra system.

2.1.2 Machine Translation Systems for Indian languages

This section summarizes the existing Machine Translation systems for Indian languages (Rao 2001).

ANGLABHARTI (1991), is a machine-aided translation system specifically designed for translating English to Indian languages. English is a SVO language while Indian languages are SOV and are relatively of free word-order. Instead of designing translators for English to each Indian language, Anglabharti uses a pseudo-interlingua approach. It analyses English only once and creates an intermediate structure called PLIL (Pseudo Lingua for Indian Languages). This is the basic translation process translating the English source language to PLIL with most of the disambiguation having been performed. The PLIL structure is then converted to each Indian language through a process of text-generation. The effort in analyzing the English sentences and translating into PLIL is estimated to be about 70% and the text-generation accounts for the rest of the 30%. Thus only with an additional 30% effort, a new English to Indian language translator can be built. The attempt has been made to 90% translation task to be done by machine and 10% left to the human post-editing. The project has been applied mailnly in the domain of public health.

Anusaaraka (1995) was developed at IIT Kanpur, and was later shifted to the Center for Applied Linguistics and Translation Studies (CALTS), Department of Humanities and Social Studies, University of Hyderabad. Of late, the Language Technology Research Center (LTRC) at IIIT Hyderabad is attempting an English-Hindi Anusaaraka MT System. The focus in Anusaaraka is not mainly on Machine Translation, but on Language access between Indian Languages. Using principles of Paninian Grammar (PG), and exploiting the close similarity of Indian languages, it essentially maps local word groups between the source and target

languages. Where there are differences between the languages, the system introduces extra notation to preserve the information of the siurce language. The project has developed Language Accesors for Punjabi, Bengali, Telugu, Kannada and Marathi into Hindi. The output generated is understandable but not grammatically correct. For example, a Bengali to Hindi Anusaaraka can take a Bengali text and produce output in Hindi which can be understood by the user but will not be grammatically perfect. The system has mainly been applied for children‘s stories.

Anubharati (1995), used EBMT paradigm for Hindi to English translation. The translation is obtained by matching the input sentences with the minimum distance example sentences. The system stored the examples in generalized form to contain the category/class information to a great extent. This made the example-base smaller in size and its further processing partitioning reduces the search space. This approach works more efficiently for similar languages such as among Indian languages.

The Mantra (MAchiNe assisted TRAnslation tool) (1999) translates English text into Hindi in a specified domain of personal administration specifically gazette notifications pertaining to government appointments, office orders, office memorandums and circulars. It is based on the TAG formalism from University of Pennsylvania. In addition to translating the content, the system can also preserve the formatting of input word documents across the translation. The Mantra approach is general, but the lexicon/grammar has been limited to the language of the domain. This project has also been extended for Hindi-English and Hindi-Bengali language pairs and also existing English- Hindi translation has been extended to the domain of parliament proceeding summaries.

MAT (2002), a machine assisted translation system for translating English texts into Kannada, has been developed by Dr. K. Narayana Murthy at Resource Centre for Indian Language Technology Solutions, University of Hyderabad. Their approach is based on using the Universal Clause Structure Grammar (UCSG) formalism. The input sentence is parsed by

UCSG parser and outputs the number, type and inter-relationships amongst various clauses in the sentence and the word groups that take on various functional roles in clauses. Keeping this structure in mind, a suitable structure for the equivalent sentence in the target language is first developed. For each word, a suitable target language equivalent is obtained from the bilingual dictionary. The MAT System provides for incorporating syntactic and some simple kinds of semantic constraints in the bilingual dictionary. The MAT system includes morphological analyzer/generator for Kannada. Finally, the target language sentence is generated by placing the clauses and the word groups in appropriate linear order, according to the constraints of the target language grammar. Post Editing tool has been provided for editing the translated text. MAT System 1.0 had shown about 40-60% of fully automatic accurate translations. It has been applied to the domain of government circulars, and funded by the Karnataka government.

An English–Hindi Translation System (2002) with special reference to weather narration domain has been designed and developed by Lata Gore et. al. The system is based on transfer based translation approach. MT system transfers the source sentence to the target sentence with the help of different grammatical rules and also a bilingual dictionary. The translation module consists of sub modules like Pre-processing of input sentence, English tree generator, post-processing of English tree, generation of Hindi tree, Post-processing of Hindi tree and generating output. The translation system gives domain specific translation with satisfactory results. By modifying the database it can be extended to other domains.

VAASAANUBAADA (2002), an Automatic Machine Translation of Bilingual Bengali-Assamese News Texts using Example-Based Machine Translation technique, has been developed by Kommaluri Vijayanand et. al. It involves Machine Translation of bilingual texts at sentence level. In addition, it also includes preprocessing and post-processing tasks. The bilingual corpus has been constructed and aligned manually by feeding the

real examples using pseudo code. The longer input sentence is fragmented at punctuations, which results in high quality translation. Backtracking is used when the exact match is not found at the sentence/fragment level, leading to further fragmentation of the sentence. The results when tested by authors are fascinating with quality translation.

ANGLABHARTI-II (2004) addressed many of the shortcomings of the earlier architecture. It uses a generalized example-base (GEB) for hybridization besides a raw example-base (REB). During the development phase, when it is found that the modification in the rule-base is difficult and may result in unpredictable results, the example-base is grown interactively by augmenting it. At the time of actual usage, the system first attempts a match in REB and GEB before invoking the rule-base. In AnglaBharti-II, provisions were made for automated pre-editing & paraphrasing, generalized & conditional multi-word expressions, recognition of named-entities. It incorporated an error-analysis module and statistical language-model for automated post-editing. The purpose of automatic pre-editing module is to transform/paraphrase the input sentence to a form which is more easily translatable. Automated pre-editing may even fragment an input sentence if the fragments are easily translatable and positioned in the final translation. Such fragmentation may be triggered by in case of a failure of translation by the 'failure analysis' module. The failure analysis consists of heuristics on speculating what might have gone wrong. The entire system is pipelined with various sub-modules. All these have contributed significantly to greater accuracy and robustness to the system.

The MaTra system (2004), a tool for human aided Machine Translation from English to Indian languages currently Hindi, has been developed by the Natural Language group of the Knowledge Based Computer Systems (KBCS) division at the National Centre for Software Technology (NCST), Mumbai (currently CDAC, Mumbai). The system has been applied mainly in the domain of news, annual reports and technical phrases. This system

used transfer approach using a frame-like structured representation. The system used rule-bases and heuristics to resolve ambiguities to the extent possible. It has a text categorization component at the front, which determines the type of news story (political, terrorism, economic, etc.) before operating on the given story. Depending on the type of news, it uses an appropriate dictionary. It requires considerable human assistance in analyzing the input. Another novel component of the system is that given a complex English sentence, it breaks it up into simpler sentences, which are then analyzed and used to generate Hindi. The system can work in a fully automatic mode and produce rough translations for end users, but is primarily meant for translators, editors and content providers.

ANUBHARTI-II (2004) has been generalized to cater to Hindi as source language for translation to any other Indian language, The system used hybrid Example-based Machine Translation approach which is a combination of example-based approach and traditional rule-based approach. The example-based approaches emulate human-learning process for storing knowledge from past experiences to use it in future. It also uses a shallow parsing of Hindi for chunking and phrasal analysis. The input Hindi sentence is converted into a standardization form to take care of word-order variations. The standardized Hindi sentences are matched with a top level standardized example-base. In case no match is found then a shallow chunker is used to fragment the input sentence into units that are then matched with a hierarchical example-base. The translated chunks are positioned by matching with sentence level example base. Human post-editing is performed primarily to introduce determiners that are either not present or difficult to estimate in Hindi.

Shakti (2004), is a Machine Translation system from English to any Indian language currently being developed at Language Technologies Research Centre, IIIT-Hyderabad. It has already produced output from English to three different Indian languages – Hindi, Marathi, and Telugu. It combines rule based approach with statistical approach. The rules are mostly linguistic in nature and the statistical approach tries to infer or use

linguistic information. Although the system accommodates multiple approaches, the backbone of the system is linguistic analysis. The system consists of 69 different modules. About 9 modules are used for analyzing the source language (English), 24 modules are used for performing bilingual tasks such as substituting target language roots and reordering etc., and the remaining modules are used for generating target language. The overall system architecture is kept extremely simple. All modules operate on a stream of data whose format is Shakti standard format (SSF).

Shiva (2004), is an example based Machine Translation system from English to Hindi developed at IIIT Hyderabad.

English-Telugu Machine Translation System has been developed jointly at CALTS with IIIT, Hyderabad, Telugu University, Hyderabad and Osmania University, Hyderabad. This system uses English-Telugu lexicon consisting of 42,000 words. A word form synthesizer for Telugu is developed and incorporated in the system. It handles English sentences of a variety of complexity.

Telugu-Tamil Machine Translation System has also been developed at CALTS using the available resources here. This system uses the Telugu Morphological analyzer and Tamil generator developed at CALTS. The backbone of the system is Telugu-Tamil dictionary developed as part of MAT Lexica. It also used verb sense disambiguator based on verbs argument structure.

ANUBAAD (2004) , an example based Machine Translation system for translating news headlines from English to Bengali, has been developed by Sivaji Bandyopadhyay at Jadavpur University Kolkata. During translation, the input headline is initially searched in the direct example base for an exact match. If a match is obtained, the Bengali headline from the example base is produced as output. If there is no match, the headline is tagged and the tagged headline is searched in the Generalized Tagged Example base. If a match is obtained, the output Bengali headline is to be

generated after appropriate synthesis. If a match is not found, the Phrasal example base will be used to generate the target translation. If the headline still cannot be translated, the heuristic translation strategy applied is - translation of the individual words or terms in their order of appearance in the input headline will generate the translation of the input headline. Appropriate dictionaries have been consulted for translation of the news headline.

Hinglish (2004) , a Machine Translation system for pure (standard) Hindi to pure English forms developed by R. Mahesh K. Sinha and Anil Thakur. It had been implemented by incorporating additional layer to the existing English to Hindi translation (AnglaBharti-II) and Hindi to English translation (AnuBharti-II) systems developed by Sinha. The system claimed to be produced satisfactory acceptable results in more than 90% of the cases. Only in case of polysemous verbs, due to a very shallow grammatical analysis used in the process, the system is unable to resolve their meaning.

Tamil-Hindi Machine-Aided Translation system has been developed by Prof. C.N. Krishnan at AU-KBC Research Centre, MIT Campus, Anna University Chennai. This system is based on Anusaaraka Machine Translation System architecture. It uses a lexical level translation and has 80-85% coverage. Stand-alone, API, and Web-based on-line versions have been developed. Tamil morphological analyser and Tamil-Hindi bilingual dictionary (~ 36k) are the by products of this system. They also developed a prototype of English - Tamil MAT system. It includes exhaustive syntactical analysis. Currently, it has limited vocabulary (100-150) and small set of Transfer rules.

AnglaHindi (2003) , a pseudo –interlingual rule-based English to Hindi Machine-Aided Translation System, developed by Sinha et. al. at IIIT, Kanpur. It is a derivative of AnglaBharti MT System for English to Indian languages. AnglaHindi besides using all the modules of AnglaBharti, also makes use of an abstracted example-base for translating frequently encountered noun phrases and verb phrasals. The system generates

approximately 90% acceptable translation in case of simple, complex and compound sentences upto a length of 20 words.

IBM-English-Hindi Machine Translation System has been initially developed by IBM India Research Lab at New Delhi with EBMT approach. Now, the approach has been changed to statistical Machine Translation between English and Indian languages.

English to {Hindi, Kannada, Tamil} and Kannada to Tamil Language-Pair Example Based Machine Translation (2006) has been developed by Prashanth Balajapally. It is based on a bilingual dictionary comprising of sentence-dictionary, phrases-dictionary, words-dictionary and phonetic-dictionary and is used for the Machine Translation. Each of the above dictionaries contains parallel corpora of sentences, phrases and words, and phonetic mappings of words in their respective files. Example Based Machine Translation (EBMT) has a set of 75000 most commonly spoken sentences that are originally available in English. These sentences have been manually translated into three of the target Indian languages, namely Hindi, Kannada and Tamil.

Google Translate (2007), is based on statistical Machine Translation approach, and more specifically, on research by Franz-Josef Och. Before using statistical approach, Google translate was using SYSTRAN for its translation till 2007. Currently, it is providing the facility of translation among 51 language pairs. It includes only one Indian language Hindi. The accuracy of translation is good enough to understand the translated text. [Internet Source: http://translate.google.com/]

Punjabi to Hindi Machine Translation System (2007) has been developed by Gurpreet Singh Joshan et. al. at Punjabi University Patiala. This system is based on direct word-to-word translation approach. This system consists of modules like pre-processing, word-to-word translation using Punjabi-Hindi lexicon, morphological analysis, word sense

disambiguation, transliteration and post processing. The system has reported 92.8% accuracy.

Sampark: Machine Translation System among Indian languages (2009), developed by the Consortium of Institutions. Consortium of institutions include IIIT Hyderabad, University of Hyderabad, CDAC(Noida,Pune), Anna University, KBC, Chennai, IIT Kharagpur, IIT Kanpur, IISc Bangalore, IIIT Alahabad, Tamil University, Jadavpur University. Currently experimental systems have been released namely {Punjabi,Urdu, Tamil, Marathi} to Hindi and Tamil-Hindi Machine Translation systems. The accuracy of the translation is not up to the mark.[Internet Source:http://sampark.iiit.ac.in]

Yahoo! Bable Fish (2008), developed by AltaVista, is a web-based application on Yahoo! that machine translates text or web pages from one of several languages into another. The translation technology for Babel Fish is provided by SYSTRAN. It translates among English, Simplified Chinese, Traditional Chinese, Dutch, French, German, Greek, Italian, Japanese, Korean, Portuguese, Russian, Swedish, and Spanish. [Internet Source: http://babelfish.yahoo.com/]

Microsoft Bing Translator (2009) is a service provided by Microsoft as part of its Bing services which allow users to translate texts or entire web pages into different languages. All translation pairs are powered by Microsoft Translation (previously Systran), developed by Microsoft Research, as its backend translation software. The translation service is also using statistical Machine Translation strategy to some extent [Internet Source: http://www.microsofttranslator.com/]

Bengali to Hindi Machine Translation System (2009) is a hybrid Machine Translation system, developed at IIT Kharagpur. This system uses multi-engine Machine Translation approach. It is based on the unfactored Moses SMT system with Giza++ (Josef,2000) derived phrase table as a central element. This system uses dictionary consisting of

15,000 parallel sysnets, Gazeteer list consisting of 50,000 parallel name list, monolingual corpus of 500K words both from source and target languages, suffix list of 100 Bengali linguistic suffixes. The BLUE score obtained during system evaluation is 0.2318.

2.2 History of Machine Translation

Looking at the history of machine translation (Hutchins, 1986, 1994, 2005; Solcum 1985), it will be surprised to know that the first idea of machine translation a universal language, with equivalent ideas in different tongues sharing one symbol proposed by René Descartes in 17th century in order to overcome the barriers in communication due to language. But it was only in the 20th century, the first concrete proposals to machine translation have been made by George Artsrouni, a French-Armenian and by Petr Smirnov-Troyanskii, a Russian, independently in 1933.

Artsrouni designed a storage device on paper tape which could be used to find the equivalent of any word in another language; a prototype was apparently demonstrated in 1937. Troyanskii envisioned the three stages of mechanical translation: first, the logical analysis of words in the source language into their base forms and syntactic functions by an editor who knows only the source language; second, the machine transforms these base forms and syntactic functions into its equivalent in the target language; third, the output of the machine is transformed into word forms in the target language manually by an editor who knows the target language. He also envisioned both the bilingual and multilingual translation. Even though, in his idea the role of machine lies only in the second stage, he said that the logical analysis will be also automated, in the years to come.

It was in January 1954, the first public demonstration of machine translation was done, in Georgetown university as a result of the project 'The Georgetown experiment' of 1954 by the Georgetown university in Collaboration with IBM. In this experiment, a carefully selected sample of 49 Russian sentences was translated into English, using a very restricted vocabulary of 250 words and just 6 grammar rules. The experiment was a great success and ushered in an era of substantial funding for machine-translation research. The authors claimed that within three to five years,

machine translation would be a solved problem. The decade of 1956 – 1965 was considered as a decade of high expectations and also the decade which destroyed the false belief that the problem of machine translation could be solved in just a few years. This was mainly because most of the people in this area of research, aimed at developing immediate systems for translation without considering the various issues in machine translation. But it was too late when they understood that it was impossible to produce translation systems over a short span of time. The problem of disillusion increased as the linguistic complexity gets more and more apparent.

As the progress shown by the researchers was very much slower and also as it failed to fulfill the expectations of the governments and companies, who funded their research, the government sponsors of MT in the United States formed the Automatic Language Processing Advisory Committee (ALPAC) to examine the prospects in 1964. It concluded in its famous 1966 report that machine translation was slower, less accurate and twice as expensive as human translation and that there is no immediate or predictable prospect of useful machine translation. It saw no need for further investment in machine translation research; instead it recommended the development of machine aids for translators, such as automatic dictionaries, and continued support of basic research in computational linguistics. The ALPAC report was widely condemned as narrow, biased and short sighted. It is true that it failed to recognize, for example, that revision of manually produced translations is essential for high quality, and it was unfair to criticize machine translation for needing to post-edit output. It may also have misjudged the economics of computer-based translation, but large-scale support of current approaches could not continue. The influence of the ALPAC report was profound. It brought a virtual end to machine translation research in the USA for over a decade and MT was for many years perceived as a complete failure.

After the ALPAC report, as United States concentrated mainly on translating the Russian's scientific and technical materials and as the need for machine translation has increased in Europe and Canada, the focus of machine translation research switched from the United States to Europe and Canada. The decade of 1967 – 1976, was considered to be a quite decade in the history of machine translation. In the 1980s, machine translation research diversified in all directions

and many commercial translation systems came into market. Research after the mid-1970s had three main strands: first, the development of advanced transfer systems building upon experience with earlier interlingua systems; secondly, the development of new kinds of interlingua systems; and thirdly, the investigation of techniques and approaches from Artificial Intelligence.

In the latter part of the 1980s developments in syntactic theory, in particular unification grammar, Lexical Functional Grammar and Government Binding theory, began to attract researchers, although their principal impact was to come in the 1990s. At the time, many observers believed that the most likely source of techniques for improving machine translation quality lay in research on natural language processing within the context of artificial intelligence.

The dominant framework of machine translation research until the end of the 1980s was based on essentially linguistic rules of various kinds: rules for syntactic analysis, lexical rules, and rules for lexical transfer, rules for syntactic generation, rules for morphology, etc. The rule-based approach was most obvious in the dominant transfer systems such as Ariane, Metal, SUSY, Mu and Eurotra, but it was at the basis of all the various interlingua systems - both those which were essentially linguistics-oriented such as DLT and Rosetta, and those which were knowledge-based.

Since 1989, however, the dominance of the rule-based approach has been broken by the emergence of new methods and strategies which are now loosely called ‘corpus-based’ methods. Firstly, a group from IBM published in 1988 the results of experiments on a system based purely on statistical methods. The effectiveness of the method was a considerable surprise to many researchers and has inspired others to experiment with statistical methods of various kinds in subsequent years. Secondly, at the very same time certain Japanese groups began to publish preliminary results using methods based on corpora of translation examples, i.e. using the approach now generally called ‘example-based’ translation. For both approaches the principal feature is that no syntactic or semantic rules are used in the analysis of texts or in the selection of lexical equivalents.

The most dramatic development has been the revival of the statistics-based approach to machine translation in the Candide project at IBM. Statistical methods were common in the earliest period of machine translation research, in the 1960s, but the results had been generally disappointing. With the success of newer stochastic techniques in speech recognition, the IBM team at Yorktown Heights began to look again at their application to machine translation. The distinctive feature of Candide is that statistical methods are used as virtually the sole means of analysis and generation; no linguistic rules are applied. The IBM research is based on the vast corpus of French and English texts contained in the reports of Canadian parliamentary debates i.e., the Canadian Hansard. The essence of the method is first to align phrases, word groups and individual words of the parallel texts, and then to calculate the probabilities that any one word in a sentence of one language corresponds to a word or words in the translated sentence with which it is aligned in the other language.

Most researchers were surprised, particularly those involved in rule-based approaches, by the results which were so acceptable: almost half the phrases translated either matched exactly the translations in the corpus, or expressed the same sense in slightly different words, or offered other equally legitimate translations. Obviously, the researchers have sought to improve these results, and the IBM group proposes to introduce more sophisticated statistical methods, but they also intend to make use of some minimal linguistic information, e.g. the treatment of all morphological variants of a verb as a single word, and the use of syntactic transformations to bring source structures closer to those of the target language.

The second major corpus-based approach - benefiting likewise from improved rapid access to large databanks of text corpora is what is known as the example-based or memory-based approach. Although first proposed in 1984 by Makoto Nagao, it was only towards the end of the 1980s that experiments began, initially in some Japanese groups and during the DLT project. The underlying hypothesis is that translation often involves the finding or recalling of analogous examples, i.e. how a particular expression or some similar phrase has been translated before. The example-based approach is founded on processes of extracting and selecting equivalent phrases or word groups from a databank of parallel bilingual texts, which

have been aligned either by statistical methods similar to those used by the IBM group or by more traditional rule-based morphological and syntactic methods of analysis. For calculating matches, some MT groups use semantic methods, e.g. a semantic network or a hierarchy of domain terms. Other groups use statistical information about lexical frequencies in the target language. The main advantage of the approach is that since the texts have been extracted from databanks of actual translations produced by professional translators there is an assurance that the results will be accurate and idiomatic.

Although the main innovation since 1990 has been the growth of corpus-based approaches, rule-based research continues in both transfer and interlingua systems. For example, a number of researchers involved in Eurotra have continued to work on the theoretical approach developed, e.g. the CAT2 system at Saarbrücken, and one of the fruits of Eurotra research has been the PaTrans transfer-based system developed in Denmark for Danish/English translation of patents.

One consequence of developments in example-based methods has been that much greater attention is now paid to questions of generating good quality texts in target languages than in previous periods of machine translation activity when it was commonly assumed that the most difficult problems concerned analysis, disambiguation and the identification of the antecedents of pronouns. In part, the impetus for this research has come from the need to provide natural language output from databases, i.e. translation from the artificial and constrained language used to represent database contents into the natural language of database users. Some machine translation teams have researched multilingual generation.

The use of machine translation accelerated in the 1990s. The increase has been most marked in commercial agencies, government services and multinational companies, where translations are produced on a large scale, primarily of technical documentation. This is the major market for the mainframe systems: Systran, Logos, METAL, and ATLAS. All have installations where translations are being produced in large volumes. Indeed, it has been estimated that in 1993 over 300 million words a year were translated by such services: for example, one Logos operation alone at

Lexi-Tech, Canada was translating annually more than 25 million words of technical manuals.

2.3 Need for MT

Machine Translation system are needed to translate literary works which from any language into native languages. The literary work is fed to the MT system and translation is done. Such MT systems can break the language barriers by making available work rich sources of literature available to people across the world. MT also overcomes the technological barriers. Most of the information available is in English which is understood by only 3% of the population. This has lead to digital divide in which only small section of society can understand the content presented in digital format. MT can help in this regard to overcome the digital divide.

2.4 Problems in MT

There are several structural and stylistic differences among languages, which make automatic translation a difficult task. Some of these issues are as follows.

- **Word order**

  Word order in languages differs. Some classification can be done by naming the typical order of subject (S), verb (V) and object (O) in a sentence. Some languages have word orders as SOV. The target language may have a different word order. In such cases, word to word translation is difficult. For example, English language has SVO and Hindi language has SOV sentence structure.

- **Word sense**

  The same word may have different senses when being translated to another language. The selection of right word specific to the context is important.

- **Pronoun Resolution**

The problem of not resolving the pronominal references is important for machine translation. Unresolved references can lead to incorrect translation.

- **Idioms**

  An idiomatic expression may convey a different meaning, that what is evident from its words. For example, an idiom in English language *'Jack of all trades'*, would not convey the intend meaning when translated into Tamil language.

- **Ambiguity**

  In computational linguistics, Word Sense disambiguation (WSD) is an open problem of natural language processing, which governs the process of identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings.

2.5 Types of Machine Translation Systems

The following are four types of Machine Translation (MT) systems:

- **MT for Watcher (MT-W)**

MT for watchers is intended for readers who wanted to gain access to some information written in foreign language who are also prepared to accept possible bad 'rough' translation rather than nothing. This was the type of MT envisaged by the pioneers. This came in with the need to translate military technological documents.

- **MT for revisers (MT-R)**

MT for revisers aims at producing raw translation automatically with a quality comparable to that of the first drafts produced by human. The translation output can be considered only as brush-up so that the professional translator can be freed from that boring and time consuming task.

- **MT for translators (MT-T)**

MT for translator's aims at helping human translators do their job by providing online dictionaries, thesaurus and translation memory. This type of machine translation system is usually incorporated into the translation work stations and the PC based translation tools.

- **MT for Authors (MT-A)**

  MT for authors aims at authors wanting to have their texts translated into one or several languages and accepting to write under control of the system or to help the system disambiguate the utterance so that satisfactory translation can be obtained without any revision.

2.6 Different Approaches used for Machine Translation

There are a number of approaches used for MT. But mainly three approaches are used. These are discussed below:

- Linguistic or Rule Based Approaches
  - Direct Approach
  - Interlingua Approach
  - Transfer Approach
- Non-Linguistic Approaches
  - Dictionary Based Approach
  - Corpus Based Approach
    - Example Based Approach
    - Statistical Approach
- Hybrid Approach

2.6.1 Linguistic or Rule-Based Approaches

Rule based approaches requires a lot of linguistic knowledge during the translation and so it uses grammar rules and computer programs which will be helpful in analysing the text for determining grammatical information and features for each and every word in the source language, translating it by replacing each word by lexicon or word that have the same context in the target language. Rule based

approach is the principal methodology that was developed in machine translation. Linguistic knowledge will be required in order to write the rules for this type of approaches. These rules will play a vital role during the different levels of translation.

The benefit of rule based machine translation method is that it can intensely examine the sentence at its syntax and semantic levels. There are complications in this method such as prerequisite of vast linguistic knowledge and very huge number of rules is needed in order to cover all the features in a language.

The three different approaches that require linguistic knowledge are as follows:

1. Direct MT
2. Interlingua MT
3. Transfer MT

2.6.1.1 Direct MT System:

Direct MT form of MT is the most basic one. It translates the individual words in a sentence from one language to another using a two-way dictionary. It makes use of very simple grammar rules. These systems are based upon the principle that as MT system should do as little work as possible. Direct MT systems take a monolithic approach towards development, *i.e.*, they consider all the details of one language pair.

Direct MT has following characteristics:

- Little analysis of source language
- No parsing
- Reliance on large two-way dictionary

The general procedure for direct translation systems can be summarized as shown in Figure 2.1. The direct MT system starts with morphological analysis. Morphological analysis removes morphological inflections from the words to get the root word from the source language words. The next step in direct MT system is

bilingual dictionary lookup. A bilingual dictionary is looked up to get the target-language words corresponding to the source-language words. The last step in direct MT system is syntactic rearrangement. In syntactic rearrangement, the word order is changed to that which best matches the word order of the target language.

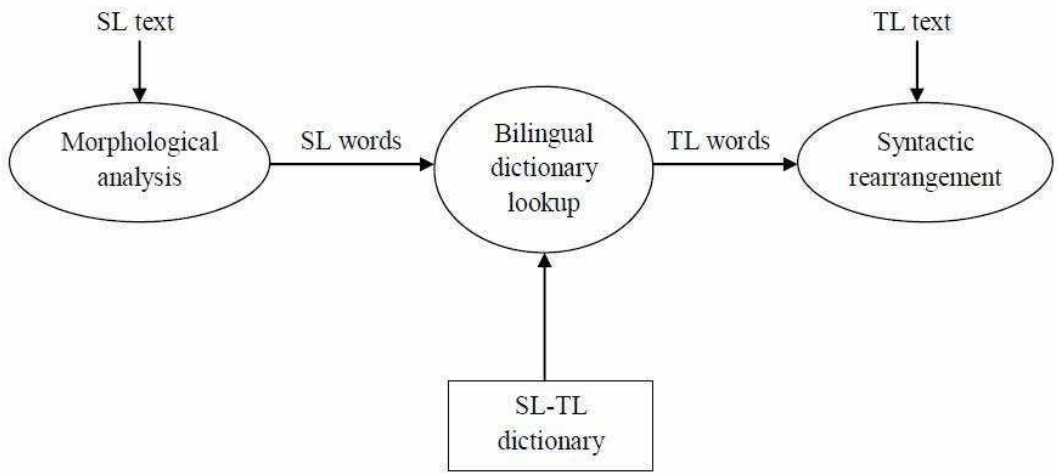


Figure 2.1: Direct Machine Translation

Direct Machine Translation works well with languages which have same default sentence structure.

Advantages of Direct MT

The Direct MT systems have below mentioned advantages.

- Translation is usually comprehended by the reader with little effort.

Disadvantage of Direct MT

The Direct MT systems have following disadvantages.

- Direct MT involves only lexical analysis. It does not consider structure and relationships between words.
- Direct MT systems are developed for a specific language pair and cannot be adapted for different language pairs.

- Direct MT systems can be quite expensive, for multilingual scenarios.
- Some of the source text meaning can be lost in the translation.

2.6.1.2 Interlingua Machine Translation

Inter is a sub version of Direct Machine Translation. The Interlingua Machine Translation converts words into a universal language that is created for the MT simply to translate it to more than one language. Figure 2.2 shows how different languages A,
B, C, D can be translated through this system.

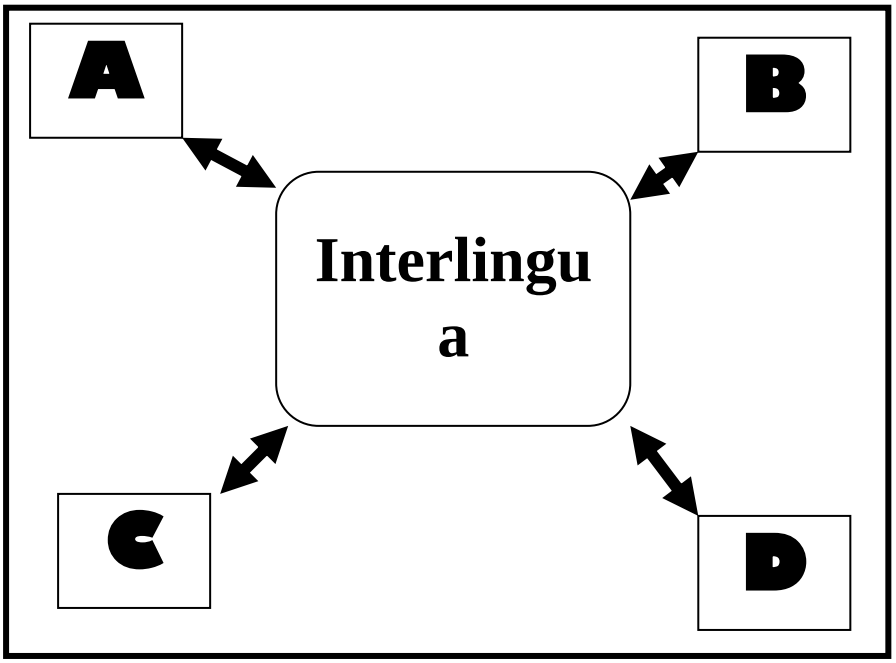


Figure 2.2: Interlingua language system

Advantages of Interlingua Machine Translation

Interlingua MT systems have below mentioned advantages.

- It gives a meaning-based representation and can be used in applications like information retrieval.

- An Interlingua system has to resolve all the ambiguities so that translation to any language can take place from the Interlingua representation.
- The system is more practical when several languages are to be interpreted since it only needs to translate it from the source language. Figure 2.3 shows how language A can be translated into several languages.
- For specific domains, Interlingua approach can be used successfully.

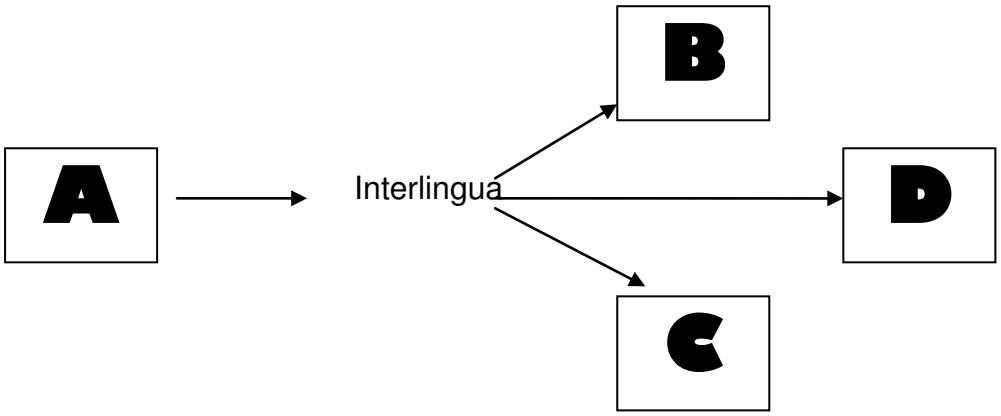


Figure 2.3: Multilingual MT system with Interlingua approach

Disadvantage of Interlingua Machine Translation

Interlingua MT systems have following disadvantages.

- Time efficiency of this system is lower than the Direct Machine Translation system.
- Major problem lies in defining a universal abstract (Interlingua) representation which preserves the meaning of a sentence.
- Defining a vocabulary for a universal Interlingua is extremely difficult as different languages conceptualize the world in different ways.
- There may be many concepts in a language or culture which lack representation in another language.

2.6.1.3 Transfer based MT

In this translation system, a database of translation rules is used to translate text from source to target language. Whenever a sentence matches one of the rules, or examples, it is translated directly using a dictionary. It goes from the source language to a morphological and syntactic analysis to produce a sort of Interlingua on the base forms of the source language, from this it translates it to the base forms of the target language and from there a better translation is made to create the final step in the translation. The steps which are performed are shown in Figure 2.4.

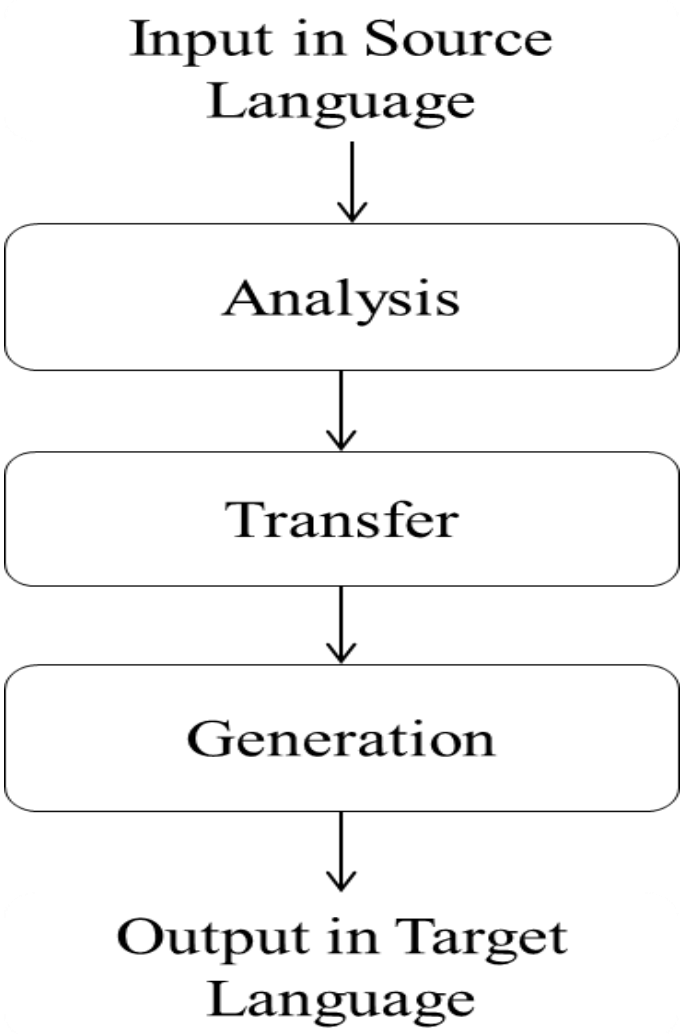


Figure 2.4: Description of Transfer-Based Machine Translation

The major modules in transfer based MT is as follows.

Analysis: Analysis phase is used to produce source language structure.
Transfer: Transfer phase is used to transfer source language representation to a target level representation.
Generation: Generation phase is used to generate target language text using target level structure.

Advantages of Transfer-Based MT

Transfer-based approach has following advantages.

- It has a modular structure.
- The system easily handles ambiguities that carry over from one language to another.

Disadvantage of Transfer-Based MT

Transfer-based MT systems have following disadvantages.

- Some of the source text meaning can be lost in the translation.

2.6.2 Non-Linguistic Approaches

The non-linguistic approaches are those which don‘t require any linguistic knowledge explicitly to translate texts in the source language to target language. The only resource required by this type of approaches is data either the dictionaries for the dictionary based approach or bilingual and monolingual corpus for the empirical or corpus based approaches.

2.6.2.1 Dictionary Based Approach

The dictionary based approach to machine translation uses s dictionary for the language pair to translate the texts in the source language to target language. In this approach, word level translations will be done. This dictionary based approach can either be preceded by some pre-processing stages to analyse the morphological

information and lemmatize the word to be retrieved from the dictionary. This kind of approach can be used to translate the phrases in a sentence and found to be least useful in translating a full sentence. This approach will be very useful in accelerating the human translation, by providing meaningful word translations and limiting the work of humans to correcting the syntax and grammar of the sentence.

2.6.2.2 Empirical or Corpus Based Approaches

The corpus based approaches don‘t require any explicit linguistic knowledge to translate the sentence. But a bilingual corpus of the language pair and the monolingual corpus of the target language are required to train the system to translate a sentence. This approach has driven lots of interest world-wide, from late 1980s till now.

2.6.2.2.1 Example Based Approach

This approach to machine translation is a technique that is mainly based how human beings interpret and solve the problems. That is, normally the humans split the problem into sub problems, solve each of the sub problems with the idea of how they solved this type of similar problems in the past and integrate them to solve the problem in whole. This approach needs a huge bilingual corpus of the language pair among which translation has to be performed. Figure 2.5 shows the block diagram of example-based approach.

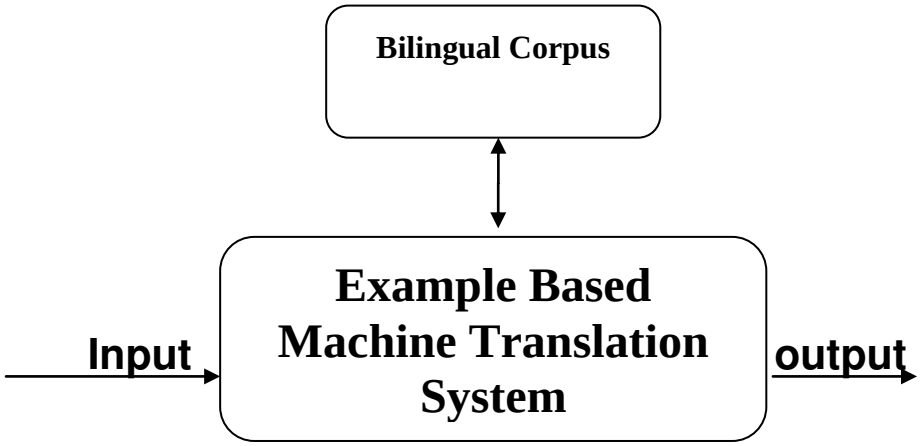


Fig. 2.5 Block diagram of example based machine translation system

In order to get a clear idea of this approach, consider the following sentence, “He bought a book”. Assuming that we are using a corpus that contains the following two sentence pairs:

English	**Tamil**
He bought a book	அவன் ஒரு புத்தகம் வாங்கினான்
He has a car	அவனுக்கு ஒரு கார் இருக்கிறது

The parts of the sentence to be translated will be matched with these two sentences in the corpus. Here, the part of the sentence ‘He bought‘ gets matched with the words in the first sentence pair and ‘a car‘ gets matched with the words in the second sentence pair. Therefore, the corresponding Tamil part of the matched segments of the sentences in the corpus are taken and combined appropriately. Sometimes, post-processing may be required in order to handle numbers, gender if exact words are not available in the corpus.

2.6.2.2.2 Statistical Approach

Statistical approach to machine translation generates translations using statistical methods by deriving the parameters for those methods by analysing the bilingual corpora. This approach differs from the other approaches to machine translation in many aspects. Figure 2.6 shows the simple block diagram of a statistical machine translation system.

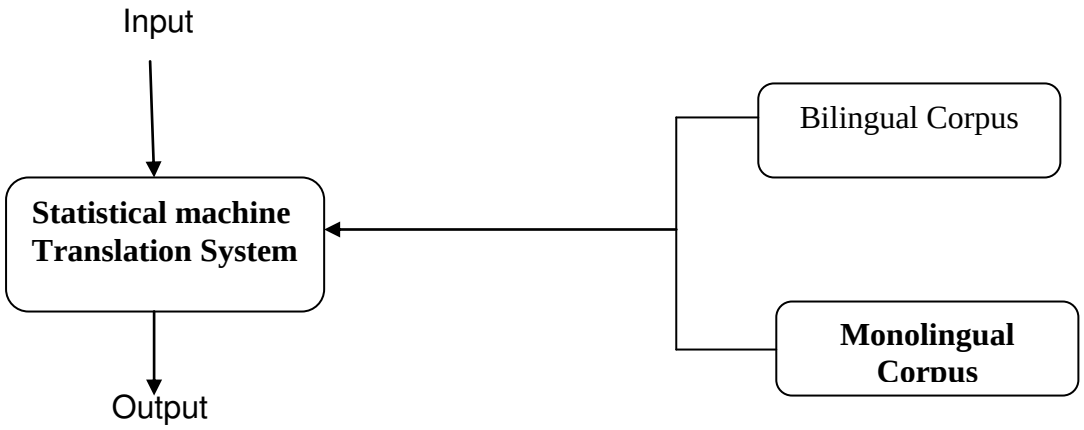


Fig. 2.6 Simple block diagram of statistical machine translation system

The advantages of statistical approach over other machine translation approaches are as follows:

- The enhanced usage of resources available for machine translation such as manually translated parallel and aligned texts of a language pair, books available in both languages and so on. That is large amount of machine readable natural language texts are available with which this approach can be applied.
- In general, statistical machine translation systems are language independent i.e., it is not designed specifically for a pair of language.
- Rule based machine translation systems are generally expensive as they employ manual creation of linguistic rules and also these systems cannot be generalised for other languages, whereas statistical systems can be generalised for any pair of languages, if bilingual corpora for that particular language pair is available.
- Translations produced by statistical systems are more natural compared to that of other systems, as it is trained from the real time texts available from bilingual corpora and also the fluency of the sentence will be guided by a monolingual corpus of the target language.

This approach makes use of translation and language models generated by analysing and determining the parameters for these models from the bilingual corpora and monolingual corpus of the target language, respectively. Even though

designing a statistical system for a particular language pair is a rapid process, the work lies on creating bilingual corpora for that particular language pair, as this was the technology behind this approach. In order obtain better translations from this approach, at least more than two million words if designing the system for a particular domain and more than this for designing a general system for translating particular language pair. Moreover, statistical machine translation requires an extensive hardware configuration to create translation models in order to reach average performance levels.

2.6.3 Hybrid Machine Translation Approach

Hybrid machine translation approach makes use of the advantages of both statistical and rule-based translation methodologies. Commercial translation systems such as Asia Online and Systran provide systems that were implemented using this approach. Hybrid machine translation approaches differ in many numbers of aspects:

- **Rule-based system with post-processing by statistical approach**: Here the rule based machine translation system produces translations for a given text in source language to text in target language. The output of this rule based system will be post-processed by a statistical system to provide better translations. Figure 2.7 shows the block diagram for this type of system.

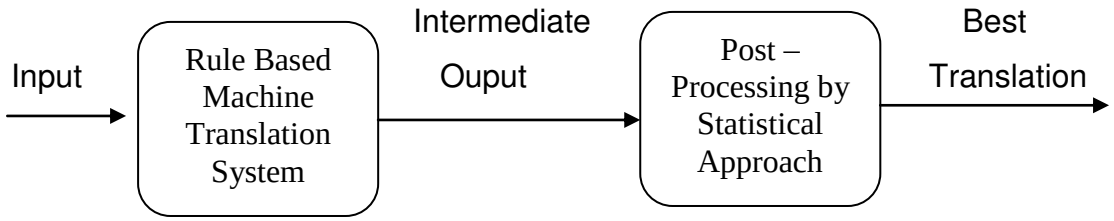


Fig. 2.7 Rule-based translation system with post-processing by statistical approach

2.7 Categories of Machine Translation System

There are three broad categories of computerized translation tools

- Fully Automated Machine Translation System
- Machine Aided Translation System

- Terminology data banks

2.7.1 Fully Automated Machine Translation System

Machine translation systems are intended to perform translation without human intervention. This does not mean that it doesn‘t need pre-processing and post-editing. However, a machine translation system is solely responsible for the complete translation process from input of the source text to output of the target text without human assistance, using special programs, comprehensive dictionaries, and collections of linguistic rules. Machine translation occupies the top range of positions on the scale of computer translation ambition.

2.7.2 Machine Aided Translation System

Machine aided translation systems generally occupy successively lower ranges on the scale of computer translation ambition.

Machine aided translation systems fall into two subgroups:

- Human-aided machine translation
- Machine-aided human translation

Human-aided machine translation refers to a system wherein the computer is responsible for producing the translation per sentence, but may interact with a human monitor at many stages along the way - for example, asking the human to disambiguate a word's part of speech or meaning, or to indicate where to attach a phrase, or to choose a translation for a word or phrase from among several candidates discovered in the system's dictionary. Machine-aided human translation refers to a system wherein the human is responsible for producing the translation per sentence, but may interact with the system in certain prescribed situations - for example, requesting assistance in searching through a local dictionary or thesaurus, accessing a remote terminology data bank, retrieving examples of the use of a word or phrase, or performing word processing functions like formatting.

The existence of a pre-processing stage is unlikely in a machine aided (human) translation system i.e., is the system does not need help, instead, it is making help available, but post-editing is frequently appropriate.

2.7.3 Terminology Data Banks

Terminology data banks are the least ambitious systems because frequent access is not made during a translation task as the translator may not be working on-line, but usually is performed prior to human translation. Indeed the data bank may not be accessible to the translator on-line at all, but may be limited to the production of printed subject-area glossaries. A terminology data banks offers access to technical terminology, but usually not to common words. The chief advantage of terminology data banks is not the fact that it is automated even with on-line access, words can be found just as quickly in a printed dictionary, but that it is up-to date: technical terminology is constantly changing and published dictionaries are essentially obsolete by the time they are available. It is also possible for terminology data banks to contain more entries because it can draw on a larger group of active contributors, its users.

2.8 Advantages of Statistical Machine Translation over Rule Based Machine Translation

Translations generated by statistical machine translation systems are better than that of traditional rule-based systems. The time duration to design a statistical machine translation system will be very much less when compared to the rule based systems. The advantages of statistical machine translation over rule based machine translation are stated below:

- Statistical machine translation system has to be trained using bilingual corpora in order to make a translation engine that translates the source language text into target language texts. In contrast, rule based machine translation system requires a great deal of knowledge apart from the corpus that only linguistic experts can generate, for example, shallow classification, syntax and semantics of all the words of source language in addition to the transfer rules between source and target languages. Rules generated are

completely reliant on one language pair involved and are not usually as studied as the classification of each separate language. Generalizing the rules is more tedious task and hence, multiple rules have to be defined for each case, particularly for languages which have different sentence structure pattern.

- Once a bilingual corpus for a particular language pair is available, more profit can be made in the translation industry by creating a statistical machine translation system for that particular language pair. In the other hand, rule based machine translation systems involves more improvement and customization costs till it touches the anticipated quality threshold. Updated rule based systems will be available at the moment when a person buys a rule based system from the market. In particular, rule based systems organisation is generally a time consuming progression including more human resources.
- Statistical systems are designed to adapt in a situation that it had not seen in the past. Whereas rule based systems have to be redesigned or retrained by the addition of new rules and words to the dictionary amid of many other things, which results in more time consumption and requires more knowledge from the linguists.
- Translations generated using statistical approach is more fluent, even though statistical systems might provide less consistency and low quality results, in case the bilingual corpus for training is too extensive for the purpose. Though rule based systems have not found the syntactic information of words suitable for analysing the source language, or does not know the word, which will prevent the finding of suitable rule.
- Statistical models and patterns are generated by statistical machine translation systems, mechanically, that includes handling exclusions regarding the rules. Concerning the rule based systems governed by the linguistic rules; they are considered as distinct case of statistical approach. However, if the rules are generalized to a large extent, they will not be able handle rule exceptions.

- Syntactic and semantic information, which are handled already in rule based systems, will also be able to handle by the statistical approach by incorporating techniques or upgrading the existing system.
- Improved quality translations will be provided by statistical systems it they are retrained with new bilingual corpus of size greater than that of previous one. Whereas, various versions of rule based systems generates more alike translations.

2.9 Applications of Machine Translation

For many years machine translation systems and tools were used principally for the production of good-quality translations: either machine translation in combination with controlled input and/or with human post-editing; or computer-based translation tools by translators. Since 1990 the situation has changed. Corporate use of machine translation with human assistance has continued to expand particularly in the area of localisation and the use of translation aids has increased particularly with the approaching of translation memories. But the main change has been the ever expanding use of unrevised machine translation output, such as online translation services provided by Babel Fish, Google, etc., applications in information extraction, document retrieval, intelligence analysis, electronic mail, and much more. The following states the various applications of machine translation briefly.

- *Traditional applications:* Machine translation has a long history – it is 60 years since Warren Weaver‘s memorandum of July 1949 launched research on the topic. For most of that history – at least 40 years – it was assumed that there were only two ways of using machine translation systems. The first was to use machine translation to produce publishable translations, generally with human editing assistance i.e., dissemination. The second was to offer the rough unedited machine translation versions to readers able to extract some idea of the content i.e., assimilation. In neither case were translators directly involved – machine translation was not seen as a computer aid for translators. The first machine translation systems operated on the traditional large-scale mainframe computers in large companies and government organizations. The

outputs of these systems were then revised (post-edited) by human translators or editors‘ familiar with both source and target languages. There was opposition from translators (particularly those with the task of post-editing) but the advantages of fast and consistent output has made large-scale machine translation cost-effective. In order to improve the quality of the raw machine translation output many large companies included methods of controlling the input language by restricting vocabulary and syntactic structures – by such means, the problems of disambiguation and alternative interpretations of structure could be minimised and the quality of the output could be improved.

- *As an aid for translators:* For most of machine translation history, translators have been wary of the impact of computers in their work. They obviously did not want to be ‗slaves‘ to mainframe machine translation output – post-editing what they could do more quickly and accurately than the machines. Many saw machine translation as a threat to their jobs – little knowing the inherent limitations of machine translation. During the 1980s and 1990s the situation changed. Translators were offered an increasing range of computer aids. First came text-related glossaries and concordances, word processing on increasingly affordable microcomputers, then terminological resources on computer databases, access to Internet resources, and finally translation memories. The idea of storing and retrieving already existing translations arose in the late 1970s and early 1980s, but did not come to fruition until the availability of large electronic textual databases and with facilitating bilingual text alignment. The first commercial translation memory systems came in the early 1990s such as Trados, Transit, Déjà Vu, WordFast, etc. All translators are now aware of their value as cost-effective aids, and they are increasingly asking for systems which go further than simple phrase and word matching – more machine translation - like facilities in other words. With this growing interest, researchers are devoting more efforts to the real computer-based needs of translators. As just two examples there are the TransSearch and TransType systems: the first a sophisticated text concordancer, the second exploiting translation memories by predicting the words a translator may select when translating a text similar to ones already translated.

- *As translators in special devices:* From the middle of the 1990s onwards, mainframe and PC translation systems have been joined by a range of other types. First should be mentioned the obvious further miniaturisation of software: the numerous commercial systems for hand-held devices. There are a bewildering variety of —pocket translatorsǁ in the marketplace. Many, such as the Ectaco range of special devices, are in effect computerized versions of the familiar phrase-book or pocket dictionary, and they are marketed primarily to the tourist and business traveller. The dictionary sizes are often quite small, and where they include phrases, they are obviously limited. However, they are sold in large numbers and for a very wide range of language pairs. As with PC systems, there is no indication of how successful in actual use they may be – it cannot be much different from the ‗successes of traditional printed phrase books. (Users may be able to ask their way to the bus station, for example, but they may not be able to understand the answer.) Recently, since early in this decade, many of these hand-held devices have included voice output of phrases, an obvious attraction for those unfamiliar with pronunciation in the target language.

- *In speech translation:* There is an increasing number of phrase-book systems offer voice output. This facility is also increasingly available for PC based translation software – it seems that Globalink in 1995 was the earliest – and it seems quite likely that it will be an additional feature for online machine translation sometime in the future. But automatic speech synthesis of text-to-text translation is not at all the same as genuine ‗speech-to-speech translation‘, the focus of research efforts in Japan (ATR), the United States (Carnegie-Mellon University), Germany (Verbmobil project) and Italy (ITC-irst, NESPOLE) for many years since the late 1980s. The research in speech translation is beset with numerous problems, not just variability of voice input but also the nature of spoken language. By contrast with written language, spoken language is colloquial, elliptical, context-dependent, interpersonal, and primarily in the form of dialogues. Machine translation has focused on well-formed, technical and scientific language and has tended to neglect informal

modes of communication. Speech translation therefore represents a radical departure from traditional machine translation. Complexities of speech translation can, however, be reduced by restricting communication to relatively narrow domains – a favourite for many researchers has been business communication, booking of hotel rooms, negotiating dates of meetings, etc. From these long-term projects no commercial systems have appeared yet. There are, however, other areas of speech translation which do have working (but not yet commercial) systems. These are communication in patient-doctor and other health consultations, communication by soldiers in military operations, and communication in the tourism domain.

- *Information retrieval:* Multilingual access to information in documentary sources (articles, conferences, monographs, etc.) was a major interest in the earliest years of machine translation, but as information retrieval (IR) became more statistics oriented and machine translation became more rule-based the reciprocal relations diminished. However, since the mid-1990s with the increasing interest in statistics-based machine translation the relations have revived, and ‗cross-language information retrieval' (CLIR) is now a vigorous area of research with strong links to machine translation: both fields are concerned with the retrieval words and phrases in foreign languages which match with words and phrases of input ‗texts' (queries in IR, source texts in machine translation), and both combine linguistic resources (dictionaries, thesaurus) and statistical techniques. There are extensions of CLIR to multilingual retrieval of images and spoken ‗documents', to retrieval of broadcast stories which are ‗similar' to a given input English text.

- *Information extraction:* Information extraction or text mining has had similar close historical links to machine translation, strengthened likewise by the growing statistical orientation of machine translation. Many commercial and government-funded international and national organisations have to scrutinize foreign-language documents for information relevant to their activities from commercial and economic to surveillance, intelligence, and espionage. The

scanning of documents received – previously an onerous human task – is now

- routinely performed automatically. Searching can focus on single texts or multilingual collections of texts, or range over selected databases e.g. via syndicated feeds or the whole Internet. The cues for relevant information include not just keywords such as ‘export’, ‘strategic’, ‗attack‘, etc. and their foreign language equivalents, but also the names of persons, companies and organisations. Since the spelling of personal names can differ markedly from one language to another, the systems need to incorporate ‘transliteration’ facilities which can convert, say, a Japanese version of a politician‘s name into its perhaps original English form. The identification of names or ‗named entities‘ and the problems of transliteration have become increasingly active fields in the last few years.

- *Other applications:*
    - *Information analysis and summarisation* is frequently the second stage after information extraction. These activities have also, until recently, been performed by human analysts. Now at least drafts can be obtained by statistical means – methods for summarisation have been researched since the 1960s. The development of working systems that combine machine translation and summarisation is apparently still something for the future.

    - The field of *question-answering* has been an active research area in artificial intelligence for many years. The aim is to retrieve answers in text form from databases in response to natural-language questions. Like summarization, this is a difficult task; but the possibility of multilingual question-answering is attracting more attention in recent years.

2.10 Summary

- A literature survey done on various machine translation systems that have been designed for Indian languages have been discussed.
- Also a literature survey on the various approaches used to handle the idioms and phrasal verbs have also been discussed.
- Followed by a theoretical background on machine translation, its history, need, different approaches such as linguistic based, non-linguistic based and hybrid has also been discussed.
- Linguistic based approaches such as transfer based, Interlingua and direct methods, non-linguistic approaches such as dictionary based, corpus based approaches have also been discussed individually in brief.
- The categories of machine translation system such as fully automated machine translation system, human aided machine translation system and machine aided human translation system have also been discussed.
- The advantages of statistical machine translation approach over rule based approach have been described.
- At the end of the chapter, the various applications of machine translation system have been discussed.

Chapter 3
Creation of Parallel Corpus

3.0 Introduction

This chapter aims to study language corpuses and parallel corpuses, their creation and their applications in machine translation. The corpus creation for Indian languages will also be discussed elaborately. McEnrey and Wilson (1996) talk in detail about corpus linguistics. Dash's (2005) contribution to corpus linguistics in the context of Indian languages is also worth mentioning.

3.1 Pre – Electronic Corpus

"Early corpus linguistics" is the term often used to describe linguistics before the advent of Chomsky. Field linguists, for example Boas (1940) who studied American-Indian languages and later linguists of the structuralist tradition all used a

corpus-based methodology. However, that does not mean that the term "corpus linguistics" was used in texts and studies from this era. Corpus was used to study language acquisition, spelling conventions and language pedagogy. The present day interpretation of corpus is different from the earlier one.

In the present era, corpus in electronic form is made use of for various purposes including NLP. Computer comes in handy to manipulate the electronic corpus. But before the advent of computer non-electronic corpuses in the hand written form were widely in use. Such non-electronic corpuses were made use of for the following tasks (Dash 2005): Corpus in dictionary making, Corpus in dialects study, Corpus for lexical study, Corpus for writing grammars, Corpus in speech study, Corpus in language pedagogy, Corpus in language acquisition and Corpus in other fields of Linguistics

3.2 Corpus in the present day context

The concept of carrying out research on written or spoken texts is not restricted to corpus linguistics. Indeed, individual texts are often used for many kinds of literary and linguistic analysis - the stylistic analysis of a poem, or a conversation analysis of a TV talk show. However, the notion of a corpus as the basis for a form of empirical linguistics is different from the examination of single texts in several fundamental ways.

Corpus linguistics is a method of carrying out linguistic analyses using huge corpuses or collections of data. As it can be used for the investigation of many kinds of linguistic questions and as it has been shown to have the potential to yield highly interesting, fundamental, and often surprising new insights about language, it has become one of the most wide-spread methods of linguistic investigation in recent years. In principle, corpus linguistics is an approach that aims to investigate linguistic phenomena through large collections of machine-readable texts. This approach is used within a number of research areas: from descriptive study of a language to the language technology and education.

In principle, any collection of more than one text can be called a corpus, (corpus being Latin for "body", hence a corpus is any body of text). But the term "corpus" when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition. The following list

describes the four main characteristics of the modern corpus (McEnery and Wilson 1996).

1. Sampling and Representativeness
2. Finite Size
3. Machine Readable Form
4. A Standard Reference

3.2.1 Sampling and Representativeness

Often in linguistics we are not merely interested in an individual text or author, but a whole variety of language. In such cases we have two options for data collection:

> We could analyse every single utterance in that variety - however, this option is impracticable except in a few cases, for example with a dead language which only has a few texts. Usually, however, analysing every utterance would be an unending and impossible task.
>
> We could construct a smaller sample of that variety. This is a more realistic option.

One of Chomsky's criticisms of the corpus approach was that language is infinite - therefore, any corpus would be skewed. In other words, some utterances would be excluded because they are rare, others which are much more common might be excluded by chance, and alternatively, extremely rare utterances might also be included several times. Although nowadays modern computer technology allows us to collect much larger corpora than those that Chomsky was thinking about, his criticisms still must be taken seriously. This does not mean that we should abandon corpus linguistics, but instead try to establish ways in which a much less biased and representative corpus may be constructed.

We are therefore interested in creating a corpus which is maximally representative of the variety under examination, that is, which provides us with an as accurate a picture as possible of the tendencies of that variety, as well as their proportions. What we are looking for is a broad range of authors and genres which,

when taken together, may be considered to "average out" and provide a reasonably accurate picture of the entire language population in which we are interested.

3.2.2 Finite Size

The term "corpus" also implies a body of text of finite size, for example, 1,000,000 words. This is not universally so - for example, at Birmingham University, John Sinclair's COBUILD team have been engaged in the construction and analysis of a **monitor corpus**. This "collection of texts" as Sinclair's team prefers to call them, is an open-ended entity - texts are constantly being added to it, so it gets bigger and bigger. Monitor corpora are of interest to lexicographers who can trawl a stream of new texts looking for the occurence of new words, or for changing meanings of old words.

Their main advantages are:

- They are not static - new texts can always be added, unlike the synchronic "snapshot" provided by finite corpora.
- Their scope - they provide for a large and broad sample of language.

Their main disadvantage is:

- They are *not* such a reliable source of **quantitative** data (as opposed to qualitative data) because they are constantly changing in size and are less rigourously sampled than finite corpora.

With the exception of monitor corpora, it should be noted that it is more often the case that a corpus consists of a finite number of words. Usually this figure is determined at the beginning of a corpus-building project. For example, the Brown Corpus contains 1,000,000 running words of text. Unlike the monitor corpus, when a corpus reaches its grand total of words, collection stops and the corpus is not increased in size. (An exception is the London-Lund corpus, which was increased in the mid-1970s to cover a wider variety of genres.)

3.2.3 Machine-readable form

Nowadays the term "corpus" nearly always implies the additional feature "machine-readable". This was not always the case as in the past the word "corpus" was only used in reference to printed text. The term *corpus* is almost synonymous with the term machine-readable corpus. Interest in the computer for the corpus linguist comes from the ability of the computer to carry out various processes, which when required of humans, ensured that they could only be described as psuedo-techniques. The type of analysis that Kading waited years for can now be achieved in a few moments on a desktop computer.

Today few corpora are available in book form - one which does exist in this way is "A Corpus of English Conversation" (Svartvik and Quirk 1980) which represents the "original" London-Lund corpus. Corpus data (not excluding context-free frequency lists) is occasionally available in other forms of media. For example, a complete key-word-in-context concordance of the LOB corpus is available on microfiche, and with spoken corpora copies of the actual recordings are sometimes available - this is the case with the Lancaster/IBM Spoken English Corpus but not with the London-Lund corpus.

Machine-readable corpora possess the following advantages over written or spoken formats:

- They can be searched and manipulated at speed. (This is something which we covered at the end of Part One).
- They can easily be enriched with extra information. (We will examine this in detail later.)
- If you haven't already done so you can now read about other characteristics of the modern corpus.

3.2.4 A standard reference

There is often a *tacit* understanding that a corpus constitutes a standard reference for the language variety that it represents. This presupposes that it will be

widely available to other researchers, which is indeed the case with many corpora - e.g. the Brown Corpus, the LOB corpus and the London-Lund corpus.

One advantage of a widely available corpus is that it provides a yardstick by which successive studies can be measured. So long as the methodology is made clear, new results on related topics can be directly compared with already published results without the need for re-computation.

Also, a standard corpus also means that a continuous base of data is being used. This implies that any variation between studies is less likely to be attributed to differences in the data and more to the adequacy of the assumptions and methodology contained in the study.

3.3 Classification of Corpus

Taking all issues under consideration we classify corpora in a broad manner in the following way (Dash 2005): genre of text, nature of data, type of text, purpose of design and nature of application.

3.3.1 Genre of Text

Based on the genre of the text the corpuses can be classified as follows:

- Written Corpus

  ex. MIT Corpus of Indian Languages contains only language data collected from various written, printed, published and electronic sources

- Speech corpus

  ex. Wellington Corpus of Spoken New Zealand English contains all formal and informal discussions, debates, previously made talks, impromptu analysis, casual and normal talks, dialogues, monologues, various types of conversations, on line dictations, instant public addressing, etc.

- Spoken corpus

  ex. London-Lund Corpus of Spoken English, a technical extension of speech corpus, contains texts of spoken language.

3.3.2 Nature of Data

Based on the nature of the data the corpuses can be classified as follows (Dash 2005):

- General corpus

  ex. British National Corpus comprises general texts belonging to different disciplines, generes, subject fields, and registers.
- Special corpus

  ex. CHILDES database is designed from text sampled in general corpus for specific variety of language, dialect and subject with emphasis on certain properties of the topic under investigation.
- Sublanguage corpus

  Sublanguge corpus consists of only one text variety of a particular language
- Sample corpus

  ex. *Zurich Corpus of English Newspapers* is one of the categories of special corpus, which are made up of small samples containing finite collection of texts chosen with great care and studied in detail.
- Literary corpus

  A special category of sample corpus is literary corpus, of which there are many kinds.

Classification criteria considered for generation of such corpus include

- author,
- genre (e.g. odes, short stories, fictions, etc.),
- period (e.g. 15th century, 18th century, etc.),
- group (e.g. Romantic poets, Augustan prose writers, Victorian novelists, etc.),
- theme (e.g. revolutionary writings, family narration, industrialisation, etc.) and
- other issues as valued parameters.

However, for some unknown reasons, corpus made from dramas and plays is usually kept separate from that of prose and poetry.

- Monitor corpus

  ex. *Bank of English* is a growing, non-finite collection of texts with scope for constant augmentation of data reflecting changes in language.

3.3.3 Types of Text

Based on the type of text the corpuses can be classified as follows (Dash 2005):

- Monolingual corpus

ex. *Bank of English* is a growing, non-finite collection of texts with scope for constant augmentation of data reflecting changes in language.

- Bilingual corpus

  ex. *MIT Bangla-Hindi Corpus* is formed when corpora of two related or non-related languages are put into one frame.

- Multilingual corpus

  ex. *Crater Corpus* contains good representative collections from more than two languages

3.3.4 Purpose of Design

Based on the purpose of design the corpuses can be classified as follows (Dash 2005):

Unannotated corpus

- *MIT Corpus of Indian Languages* represents a simple raw state of plain texts without additional linguistic or non-linguistic information.
- It has been, and is, of considerable use in language study, but utility of corpus is considerably increased by annotation

Annotated corpus

- *British National Corpus* contains tags and codes inserted from outside by designers to record some extra information (analytical marks, parts-of-speech marks, grammatical category information, etc.) into texts.
- In contrast to unannotated corpus, annotated corpus is more suitable for providing relevant information useful in various tasks for language technology including morphological processing, sentence parsing, information retrieval, word sense disambiguation, machine translation, etc..

3.3.5 Nature of Application

Based on the nature of application the corpus can be classified as follows (Dash 2005): aligned corpus, parallel corpus, reference corpus, comparable corpus and opportunistic corpus

3.3.5.1 Aligned corpus

The Canadian Hansard Corpus are a kind of bi/multi-lingual corpora where texts in one language and their translations into other language(s) are aligned, sentence by sentence, phrase by phrase or even word by word.

3.3.5.2 Parallel corpus

Chemnitz German-English/English-German Translation Corpus contains texts as well as translations in each of the languages involved allowing double-checking translation equivalents. Texts in one language and their translations into another are aligned: sentence by sentence, phrase by phrase, or even word by word. Sometimes reciprocate parallel corpora are designed where corpora containing authentic texts as well as translations in each of the languages are involved.

3.3.5.3 Reference corpus

Bank of English is designed to provide comprehensive information about a language. It aims to be large enough to represent all relevant varieties of language and characteristic vocabulary, so that it can be used as a basis for writing grammars, dictionaries, thesauruses and other reference materials. It is composed on the basis of relevant parameters agreed upon by linguistic community. It includes spoken and written, formal and informal language representing various social and situational registers. It is used as 'benchmark' for lexicons, for performance of generic tools, and language technology applications. With growing influence of internal criteria, reference corpus is used to measure deviance of special corpus.

3.3.5.4 Comparable corpus

Corpus of European Union is a collection of 'similar' texts in more than one language or variety. This kind of multilingual corpus contains texts in different languages where texts are not same in content, genre or register. These are used for comparison of different languages. It follows same composition pattern but there is no agreement on the nature of similarity, because there are few examples of comparable corpora. They are indispensable source for comparison in different languages as well as generation of bilingual and multilingual lexicons and dictionaries.

3.3.5.5 Opportunistic corpus

An opportunistic corpus stands for inexpensive collection of electronic texts that can be obtained, converted, and used free or at a very modest price; but is often unfinished and incomplete. Therefore, users are left to fill in blank spots for themselves. Their place is in situations where size and corpus access do not pose a problem. The opportunistic corpus is a virtual corpus in the sense that selection of an actual corpus (from opportunistic corpus) is up to the needs of a particular project. Monitor corpus generally considered as opportunistic corpus.

3.4 Generation of Written Corpus

There are various issues related with corpus design, development, and management. The issues of corpus development and processing may vary depending on the type of corpus and the purpose of use.

Issues related to speech corpus development differ from issues related to text corpus development. Developing a speech corpus involves issues like propose of use, selection of informants, choice of settings, manner of data-sampling, manner of data collection, size of corpus, problem of transcription, type of data encoding, management of data files, editing of input data, processing of texts, analysis of texts, etc.

Developing a written text corpus involves issues like size of corpus, representativeness, question of nativity, determination of target users, selection of time-span, selection of documents, collection of text documents (books, newspapers, magazines etc.), method of data sampling (sorting of collected materials according one's need), manner of data input (random, regular, selective, etc.), corpus sanitation (error correction omission of foreign words, quotations, dialects etc.), corpus file management, problem of copy-right etc.

3.4.1 Size of Corpus

How big will be a corpus? This points out that size is an important issue in corpus generation. It is concerned with total number of words (tokens) and different words (types) to be taken into a corpus. It also involves the decision of how many categories we like keep in corpus, how many samples of texts we put in each category, and how many words we will keep in each sample.

Although the question of size affects validity and reliability of a corpus, it is stressed that any corpus, however big, is nothing more than a minuscule sample of all speech and writing varieties produced by users of a language.

In early corpus generation era, when computer technology for procuring language data was not much advanced, it was considered that a corpus containing 1 million words or so is large enough to represent the language.

But by the mid of 1980s, computer technology went through a vast change with unprecedented growth of its storage, processing, and accessing abilities that have been instrumental in changing the concept regarding size.

Now it is believed that the bigger the size of corpus the more it is faithful in representing language. With advanced computer technology we can generate corpus of very large size containing hundreds of million of words. *Bank of English, BNC, Cobuild Corpus, Longman/Lancaster Corpus, ICE*, *ANC,* are large in size - each one containing more than 100 million words.

3.4.2 Representativeness of Texts

Within any text category, the greater the number of individual samples, the greater is the reliability of analysis of linguistic variables. The *Brown* and *LOB Corpus*, as well as *SEU* are carefully designed to that we can consider them as good representatives of the language used in America and UK. However, a simple comparison of BNC - 100 million words corpus having much more diversified structure and representative frame, with Brown, LOB, and SEU will show how these corpora are smaller in content and less diversified in structure. This easily settles empirically the issue of size and representativeness in corpus.

3.4.3 Question of Nativity

The question is whose writings should be included in corpus: the native users or non-native users?General argument is that if it is a monitor corpus then texts produced by native users should get priority over the texts of non-native users. Because, the aim of monitor corpus is to represent language, which can be considered as ‘ideal’ form for all kinds of works in linguistics and language technology. Citation of made-up examples and listing of 'ungrammatical' sentences in a monitor corpus have fairly significant effect on results of linguistic analysis of

corpus. Because, in that case, we get a lot of 'mention' rather than 'use' of words and phrases in corpus. If one of the main reasons for building a corpus is to enable us to analyse naturally occurring language, in order to see what does occur and what does not, then letting in lots of made-up example sentences and phrases will make it less fit for proposed purpose. One way of avoiding this, and many other potential problems, which are found in specialised corpus, is to apply a criterion for inclusion of texts in corpus that they should not be too technical in nature.

In case of special corpus, texts produced by non-native users are considered since the aim of a special corpus is to highlight peculiarities typical to non-native users. Here the question of representiveness of corpus is not related with the language as a whole, but with the language used by a particular class of people who have learnt and used language as their second language.

The idea is to have a corpus that includes data from which we can gather information about how a language is commonly used in various mainstreams of linguistic interactions.When we try to produce some texts and references that will provide guidance on word use, spelling, syntactic constructions, meanings, etc. most likely we would like to acquire texts of the native users.

In principle, these texts written and spoken by native users will be more directive, appropriate, and representative for enhancing ability of language understanding and use for language learners. Perhaps, this goes with rightly along the line of desire of non-native users who while learning a second language aim to achieve the efficiency of a native language user. The question of nativity becomes more complicated and case-sensitive when we find that same language is used by two different speech communities separated by geographical or political distance (e.g. British English and Indian English).

In these cases we like to recognise or generate lexical items or syntactic constructions that are common in, or typical of, a native speaker - especially those which differ from another (lexical items typical to British English vs. lexical items typical to Indian English). We also like to get into the things that are correct by the ‘rules’ of grammar and usage of Indian English, and perfectly understandable; but just not ‘right’ in rules of grammar and usage in British English. This usually betrays

the most proficient 'native' speaker of Indian English the opportunity for enlisting their languages in corpus of language used by the native speakers.

In the context when Indian people are exposed to lots of linguistic material that shows marks of being non-Indian English (Indians are exposed to lots of British English text), people who want to describe, recognise, understand, and generate Indian English will definitely ask for texts produced by native speakers of Indian English, which will highlight the linguistic traits typical to Indian English, and thus will defy all pervading influence of British English over Indian English.

3.4.4 Determination of Target Users

There are no fixed target users for general corpus. Anybody can use it for any purpose. For specialised corpus: question of target user is important. Since, each investigator or researcher has specific requirement, corpus has to be designed accordingly. A person working on developing tools for MT will require a parallel corpus rather than a general corpus. Similarly a person working on comparative studies between or more languages will require comparable corpus rather than a monitor corpus:

Target users: **Corpus**

- Descriptive linguists: General, written, and speech corpus
- NLP and LT people: General, monitor, parallel, spoken, aligned corpus
- Speech technology people: Speech corpus (text to speech, speech recognition, synthesis, processing, speech repairing, etc.)
- Lexicographers and terminologists: General, monitor, specialised, reference, opportunistic corpus etc.
- Dialogue researchers: Speech, spoken, annotated, specialised corpus
- Sociolinguistics: General, written, speech, monitor corpus
- Psycholinguistics: Specialised, speech, written corpus
- Historians: Literary, diachronic corpus
- Social scientists: General, speech, written and special corpus
- Comparative linguists: Bilingual, multilingual, parallel, comparable corpus

- MT specialists: Bilingual, multilingual, parallel, comparable, annotated corpus
- Information retrieval specialists: General, monitor, and annotated corpus
- Tagging, processing and parsing specialists: Annotated, monitor, written, spoken, general corpus
- Core-grammar designer: Comparable, bilingual, and general corpus
- Word-Sense disambiguation worker: Annotated, monitor, written, spoken, general corpus
- Teachers and students :Learner, monitor, and general corpus

3.4.5 Selection of Time-Span

Language changes with time. So determination of particular time span is required to capture features of a language within this time span. Corpus attempts to cover a particular period of time with a clear time indicator. Materials published between 1981 and 1995 are included in MIT corpus with an assumption that data will sufficiently represent the condition of present day language, and will provide information about the changes taking place within the period.

3.4.6 Selection of Texts Type

An important issue in written corpus designing is to determine if it will contain both written texts of all types. Most of the corpora incline towards written texts of standard writings. The aim of a general corpus is to identify what are central (common), as well typical (special) features of a language. Therefore, we do not require to furnish corpus with all the best pieces of contemporary writings. A measured and proportional representation will suffice. To be realistic we should include works of the mass of ordinary writers along with works of established and well-known writers.

Thus, a corpus is a collection of materials taken from different branches of human knowledge. Here writings of highly reputed authors as well as little-known writers are included with equal emphasis. All catalogues and list of publications of different publishers need to be consulted for collection of documents (books, newspapers, magazines etc.) for data collection. It is broadly heterogeneous in nature as it gathers materials from various sources and disciplines where

individuality of particular source is made obscured. Diversity is a safeguard to corpus against any kind of skewed representativeness.

The MIT Tamil corpus contains texts from Literature (20%), Fine Arts (5%), Social Science (15%), Natural Science (15%), Commerce (10%), Mass media (30%), and Translation (05%). Each category has some sub-categories. E.g., *Literature* includes novels, short stories, essays etc.; *Fine Arts* includes paintings, drawings, music, sculpture etc.; *Social Science* includes philosophy, history, education etc.; *Natural Science* includes physics, chemistry, mathematics, geography etc.; *Mass Media* includes newspapers, magazines, posters, notices, advertisements etc.; *Commerce* includes accountancy, banking etc., and *Translation* includes all the subjects translated into Tamil

3.4.7 Method of Data Sampling

Data have to be sorted from collected materials according to one's need. Sorting can be random, regular, or selective order. There are various ways for data sampling to ensure maximum representativeness of corpus. We must clearly define the kind of language we wish to study before we define sampling procedures for it. Random sampling technique saves a corpus from being skewed and unrepresentative. This standard technique is widely used in many areas of natural and social sciences.

Another way is to use complete bibliographical index. The *British National Bibliography*, and *Willing's Press Guide* are used for generation of *LOB* corpus. Another approach is to define a sampling frame. Designers of *Brown Corpus* adopted this. They used all books and periodicals published in a particular year. A written corpus may be made up of genres such as newspaper report, romantic fiction, legal statutes, scientific writing, social sciences, technical reports, and so on.

3.4.8 Method of Data Input

Data from electronic source: In this process newspapers, journals, magazines, books etc. are included if these are found in electronic form. Data from the web: This includes texts from web page, web site, and home pages. Data from e-mail: Electronic typewriting, e-mails etc. are also used as source of data. Machine

reading of text: It converts texts into machine-readable form by optical character recognition (OCR) system. Using this method, printed materials are quickly entered into corpus. Manual data input: It is done through computer keyboard. This is the best means for data collection from hand-written materials, transcriptions of spoken language, and old manuscripts. The process of data input is based on the method of sampling. We can use two pages after every ten pages are from a book. This makes a corpus best representative of data stored in physical texts. For instance, if a book has many chapters, each chapter containing different subjects written by different writers, then samples collected in this process from all chapters will be properly represented. Header File contains all physical information about the texts such as name of book, name of author(s), year of publication, edition number, name of publisher, number of pages taken for input, etc. which are required for maintaining records, and dissolving copyright problems.

It is also advantageous to keep detailed records of the materials so that documents are identified on grounds other than those, which are selected as formatives of corpus. Information whether the text is a piece of fiction or non-fiction, book, journal or newspaper, formal or informal etc. are useful for both linguistic and non-linguistic studies. At time of input, physical line of texts is maintained on screen. After a paragraph is entered, one blank line is added, and then a new paragraph is started. Texts are collected in a random sampling manner and a unique mark is put at the beginning of a new sample of text.

3.4.9 Hardware Requirement

For developing Tamil corpus they used a Personal Computer (PC) with a GIST or Transcript Card (TC), a software namely Script Processor (SP), a monitor, one conventional computer keyboard, a multilingual printer, and some floppy diskettes. Files are developed with TC installed in PC. This allows display of various Indian scripts on computer screen. Codes for various keys used in Indian characters are standardised by the Bureau of Indian Standards. With installation of this inside a PC, we can use almost the entire range of text-oriented application packages. We can also input and retrieve data in Indian language. Software also provides a choice of two operational display modes on the monitor: one in conventional English mode, and other in Indian multilingual mode.

3.4.10 Management of Corpus Files

Corpus management is a tedious task. It involves various related tasks such as holding, processing, screening, retrieving information from corpus, which require utmost care and sincerity. Once a corpus is developed and stored in computer, we need schemes for regular maintenance and augmentation. There are always some errors to be corrected, modifications to be made, and improvements to be implemented. Adaptation to new hardware and software technology and change in requirement of users are also taken care of. In addition to this, there has been constant attention to the retrieval task as well as processing and analytic tools. At present, computer technology is not so developed to execute all these works with full satisfaction. But we hope that within a few years software technology will improve to fulfil all our needs.

3.4.11. Method of Corpus Sanitation

After the input of data, the process of editing starts. Generally, four types of error occur in data entry: (a) omission or deletion of character, (b) addition or repetition of character, (c) substitution of character, and (d) transposition or displacement of character. To remove spelling errors, we need thorough checking of corpus with physical data source, and manual correction. Care has to be taken to ensure that spelling of words in corpus must resemble spelling of words used in source texts. It has to be checked if words are changed, repeated or omitted, punctuation marks are properly used, lines are properly maintained, and separate paragraphs are made for each text. Besides error correction, we have to verify omission of foreign words, quotations, dialectal forms after generation of corpus. Nativised foreign words are entered into corpus. Others are omitted. Dialectal variations are properly entered. Punctuation marks and transliterated words are faithfully reproduced.

Usually, books on natural and social sciences contain more foreign words, phrases and sentences than books of stories or fiction. Quotations from other languages, poems, songs, and dialects; mathematical expressions; chemical

formulae; geometric diagrams; tables, pictures, figures, and other symbolic representations of source texts are not entered in corpus. All kinds of processing works become easier if corpus is properly edited.

3.4.12 Problem of Copy Right

To be in the safe side we need copyright clearance from all copyright holders (publishers and/or authors, all speakers for spoken materials). Copyright laws are complicated. There is very little which is obviously right or wrong, and legal or illegal. Moreover, copyright problems differ in various countries. If one uses the material only for personal use, then there is no problem. This is fine not only for a single individual but also for a group who are working together on some areas of research and investigation. So long it is not directly used for commercial purposes, there is no problem. Using materials we can generate new tools and systems to commercialise. In that case also the copyright is not violated. The reformed generation of output provides safeguards against possible attacks from copyright holders. But in case of direct commercial work, we must have prior permission from legal copyright holders

3.5 Corpus Processing

Need for corpus processing techniques arise after accumulation large electronic corpora in many languages. People devise systems and techniques for accessing language data and extracting relevant information from corpus. These processing tools are useful for linguistic research and language technology developments. There are various corpus processing techniques (e.g., statistical analyser, concordancer, lexical collocator, key-word finder, local-word-grouper, lemmatiser, morphological processor and generator, word processor, parts-of-speech tagger, corpus annotator, parser, etc.). There are many corpus processing software available for English, French, German and similar such languages. For Indian language there are only a few. We need to design corpus-processing tools for our own languages keeping the nature of Indian languages in mind. The following is the list of text processing scheme: frequency Study, word Sorting, concordance,

lexical collocation, key word Context (KWIC), local word grouping (LWG), word processing, tagging, lemmatization, annotation and parsing.

3.5.1 Frequency Study

Linguistics is a subject, which has a long relationship with statistics and mathematics. Mathematical linguistics, computational linguistics, corpus linguistics, applied linguistics, forensic linguistics, stylometrics, etc. requires different statistical and quantitative results obtained from natural language corpus. Corpus can be subject to both *quantitative* and *qualitative* analysis. Simple descriptive statistical approach enables us to summarise the most important properties of observed data. Inferential statistical approach uses information from descriptive statistical approach to answer questions or to formulate hypothesis. Evaluative statistical approach enables to test whether hypothesis is supported by evidence in data, and how mathematical model or theoretical distribution of data relates to reality (Oakes 1998: 1).

To perform comparisons we apply multivariate statistical techniques (e.g., *Factor Analysis, Multidimensional Scaling, Cluster Analysis, Log-linear Models* etc.) to extract hidden patterns from raw frequency data obtained from corpus

3.5.2 Word Sorting

Numerical sorting is the most straightforward approach to work with quantitative data. Here items are classified according to a particular scheme, and an arithmetical count is made on the number of items within texts, which belong to each class in the scheme.Information available from simple frequency counts are rendered either in alphabetical or in numerical order. Both lists can again be arranged in ascending or descending order according to our requirement. Anyone who is studying a text will like to know how often each different item occurs in it. A frequency list of words is a set of clues to texts. By examining the list we get an idea about the structure of text and can plan an investigation accordingly. Alphabetical sorted list is used for simple general reference. A frequency list in alphabetical order plays a secondary role because it is used only when there is a need to check frequency of a particular item. However, it is useful as an object of study as it is often

helpful in formulating hypotheses to be tested, and checking assumptions that have been made before hand Kjellmer (1984).

3.5.3 Concordance

Process of concordancing is making an index to words used in corpus. It is a collection of occurrences of words, each in its own textual environment. Each word is indexed with reference to the place of each occurrence in texts. It is indispensable because it gives access to many important language patterns in texts. It provides information not accessible via intuitions. There are some concordance software available (e.g *MonoConc* for sorting and frequency, *ParaConc* for parallel texts processing), *Conc* for sorting and frequency counting), *Free Text* for processing, sorting etc.), for analysing corpus. It is most frequently used for lexicographical works. We use it to search single as well as multiword strings, words, phrases, idioms, etc. It is also used to study lexical, semantic, syntactic patterns, text patterns, genre studies, literary texts etc. (Barlow (1996). It is an excellent tool for investigating words and morphemes, which are polysemous and have multiple functions in language.

3.5.4 Lexical Collocation

Method of collocation on words helps to understand the role and position of words in texts. It helps to determine which pairs of words have a substantial collocational relation between them. It compares probabilities of two words occurring together as an event with probability that they are simply the result of chance. For each pair of words, a score is given - the higher the score the greater is the collocationality. It enables to extract multiword units from corpus to use in lexicography and technical translation. It helps to group similar words together to identify sense variations (e.g. *river bank* = landscape, but *investment in bank* = financial use.). It helps in discriminate differences in usage between words, which are similar in meaning. For instance, *strong* collocates with *motherly, showings, believer, currents, supporter*, *odour* etc. while *powerful* collocates with *tool, minority, neighbour, symbol, figure, weapon, post* etc. (Biber at al. 1998: 165)

3.5.6 Key Word In Context (KWIC)

KWIC is widely used in data processing. It helps to look up each occurrence of particular words (similar to concordance). The word under investigation appears at the centre of each line, with extra space on either side. The length of context is specified for different purposes. It shows an environment of two, three or four words on either side of the word at the centre. This pattern may vary according to one's need. At the time of analysis of words, phrases, and clauses it is agreed that additional context is needed for better understanding.

After access of a corpus by KWIC we can formulate various objectives in linguistic description and devise procedures for pursuing these objectives. KWIC helps to understand importance of context, role of associative words, actual behaviour of words in contexts, actual environment of occurrence, and if any contextual restriction is present.

3.5.7 Local Word Grouping (LWG)

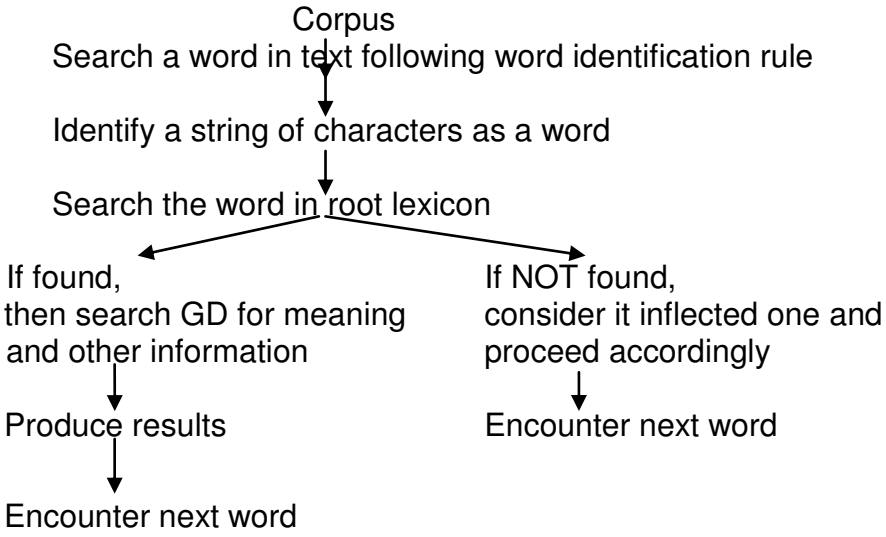
LWG is another type of text analysis, which throws light on the pattern of use of words in texts. LWG provides information for dealing with functional behaviour of constituents at the time of parsing, both in phrase and sentence level. Using LWG we find that most non-finite verbs are followed by finite verbs, while nouns are mostly followed by suffixes and post-positions in Tamil. It helps to analyse so called *verb groups* and *noun groups* from their local information. It provides clues for understanding their roles in phrases, clause, and sentences. Information from LWG helps to dissolve lexical ambiguity, which arises from local association of various lexical items. Our experience with Tamil suggests that finer shades of meaning are mostly conveyed by internal relation between constituents along with their distributions in contexts. For many compound nouns and verbs, meaning denoted by a particular association of words cannot be obtained from meanings of individual words.

3.5.8 Word Processing

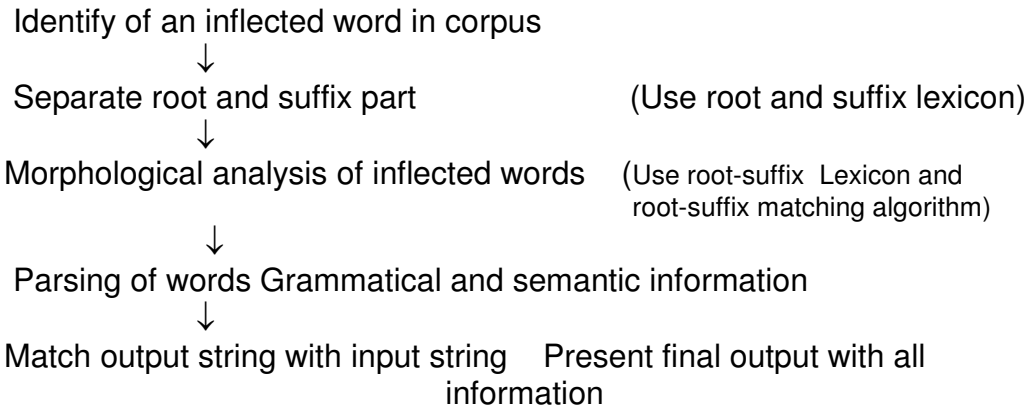
Word processing involves automatic analysis of words used in corpus. The main objective is to identify a word in a piece of text, isolate it from its contextual environment of use, analyse its morphophoemic structure, obtain its original meaning, and define its syntactic role it plays in text. Information obtained from word

processing is valuable for word sense disambiguation (WSD), dictionary making, parsing, language learning, etc. People working on native language can have better results since intuitive knowledge helps in finding out right root or suffix part form inflected words, which may be beyond the grasp of non-native users.

Processing non-inflected words


Corpus
Search a word in text following word identification rule
Identify a string of characters as a word
Search the word in root lexicon
If found,
then search GD for meaning
and other information
Produce results
Encounter next word
If NOT found,
consider it inflected one and
proceed accordingly
Encounter next word

Processing inflected words


Identify of an inflected word in corpus
Separate root and suffix part
(Use root and suffix lexicon)
Morphological analysis of inflected words
(Use root-suffix Lexicon and
root-suffix matching algorithm)
Parsing of words Grammatical and semantic information
Match output string with input string
Present final output with all
information

Processing double words

Processing double words includes compounds, reduplicated words, and detached words where constituents are separated from each other with a space in between. All detached words are multiword strings, which need to be treated in more efficient way for processing and annotation. For processing double the best method is to use delayed processing technique where processing result of one constituent is withheld until result of processing of subsequent constituent is obtained. This helps

to dissolve ambiguity at word level, since meaning of a neighbouring word helps to determine meaning of double words.

3.5.9 Tagging

Certain types of linguistic annotation, which involve attachment of special codes to words in order to indicate particular features, are often known as tagging rather than annotation, codes, which are assigned to features are known as tags.

Part-of-speech (POS) tagging

Parts-of-speech tagging scheme tags a word with its part-of-speech in a sentence. It is done at three stages: (a) pre-editing, (b) automatic tag assignment, and (c) manual post-editing. In pre-editing stage, corpus is converted to a suitable format to assigns a part-of-speech tag to each word or word combination. Because of orthographic similarity, one word may have several possible POS tags. After initial assignment of possible POS, words are manually corrected to disambiguate words in texts. An example of POS tagging is given below.

Untagged Sentence

A move to stop Mr. Gaitskell from nominating any more labour life peers is to be made at a meeting of labour MPs tomorrow.

Tagged sentence

^a_AT move_NN to_TO stop_VB \0Mr_NPT Gaitskell_NP from_IN nominating_VBG any_DTI more_AP labour_NN life_NN peers_NNS is_BEZ to_TO be_BE made_VBN at_IN a_AT meeting_NN of_IN

3.6 Parallel corpora

This section is the major concern of the present thesis. In this section the parallel corpus will be studied elaborately focusing on the creation of parallel corpus for machine translation**.**

In addition to monolingual corpora, parallel corpora have been key focus of corpus linguistics, largely because corpora of this type are important resources for translation. Parallel corpora are valuable resources on natural language processing and, in special, on the translation area. They can be used not only by translators, but also analyzed and processed by computers to learn and extract information about the languages.

Corpora in general and, particularly, parallel corpora are very important resources for tasks in the translation field like linguistic studies, information retrieval systems development or natural language processing. In order to be useful, these resources must be available in reasonable quantities, because most application methods are based on statistics. The quality of the results depends a lot on the size of the corpora, which means robust tools are needed to build and process them.

A parallel corpus contains texts in two languages. We can distinguish two main types of parallel corpus:

Comparable corpus: the texts are of the same kind and cover the same content. An example would be a corpus of articles about football from English and Danish newspapers; or legal contracts in Spanish and Greek.

Translation corpus: the texts in one language (L1) are translations of texts in the other language (L2).

Many researchers have built translation corpora in the past decade, though unfortunately most of them are not easily available. For a useful survey of parallel corpora round the world, look at Michael Barlow's parallel corpora web page (Barlow n.d).

To use a translation corpus you need a special piece of software called a **Parallel Concordancer**. With this software you can ask the computer to find all the examples of a word or phrase in L1, along with all the corresponding translated sentences in L2. Two widely-used parallel concordancers are **ParaConc** and Multiconcord.

Parallel corpora can be bilingual or multilingual, i.e. they consist of texts of two or more languages. They can be either unidirectional (e.g. an English text translated into German), bidirectional (e.g. an English text translated into German and vice versa), or multidirectional (e.g. an English text such as an EU regulation translated into German, Spanish, French, etc.).

3.6.1 Parallel Corpora Types

To discuss parallel text alignment and understand alignment problems, we will begin by pointing out some translation characteristics. We can classify translations according to the dependency between the original text and its translation:

• Type A

when the translated text will completely substitute the original text in the target language. This is the case of literary translations (where readers will choose to read only one version of them);

• Type B

when translations will coexist in time and space. This is the case of bilingual literary editions (where the reader will probably compare the texts on both languages);

• Type C

when the translations will be used for the same purpose as the original, and work in a symmetrical way. This is the case for institutional documents of the European Union and other multilingual institutions;

or classify them with respect to the translation objective:

• Pragmatic

the translated text will be used for the same communication purpose as the original;

• Stylistic

the translated text tries to maintain the original text structure and form of language;

• Semantic

the translated text tries to transmit essentially the same message.

Parallel text alignment problems are highly dependent on these classifications:

• type A translations cannot be viewed as parallel corpora. The translator often changes the order of sentences and some content8 as soon as they maintain the basic idea behind the text;

• type B translations give reasonable results on word alignment, as most specific terms from the corpora will be coherently translated between sentences;

• type C

translations are the best type of parallel corpora for alignment. As this type of parallel corpora is normally composed of institutional documents with laws and other important information, translation is done accurately, so that no ambiguities are inserted in the text, and they maintain symmetrical coherence;

Considering the automatic translation objective, stylistic and semantic translation types can have problems. Stylistic approach makes the translator look for some similar sound, sentence construction, rhythm, or rhyme. This means that the translator will change some of the text semantic in favor of the text style. The semantic approach has the advantage that the text message and semantic is maintained, but the type of language can change (as the translation will be addressed to an audience that differs significantly from the one of the original text).

3.6.2 Examples of parallel corpora

The following are a few examples of parallel corpora.

- English-German Translation Corpus
- English-Norwegian Parallel Corpus (ENPC)
- English-Swedish Parallel Corpus (ESPC)
    - cf. 'Contrastive linguistics and corpora' by S. Johansson
    - cf. The website of the English-Norwegian Parallel Corpus
    - started in 1993
    - has become an important resource for contrastive studies of English and Swedish
    - contains 64 English texts + translations, 72 Swedish texts + translations
    - contains 2.8 million words
    - contain a wide range of text types, authors, translators

    - texts have been matched as far as possible in terms of text type, subject, register
    - can therefore be used as a bidirectional parallel corpus and as a comparable corpus
    - current research: epistemic modality and adverbial connectors in English and Swedish
- The International Telecommunications Union Corpus (English-Spanish)
- The Intersect Parallel Corpus (English-French)
- The Multilingual Parallel Corpus (Danish, English, French, German, Greek, Italian, Finnish, Portuguese, Spanish, Swedish texts)

3.6.3 Applications of parallel corpora

- Parallel corpora can be used for various practical purposes. Parallel corpora can be used for many tasks, e.g. teaching, terminological studies, automatic translation or cross-language information retrieval engines:
- teaching second languages/translation didactics parallel corpora can be searched by translation students to find translation samples, gather common errors done, and learn translation techniques. It can also be used in the process of learning a second language.
- By reading parallel texts, the student can try to understand the translated sentence and mentally align concepts and structures with the original one;
- terminology studies parallel corpora can be mined to bootstrap or enrich multilingual terminology dictionaries or thesaurus. In fact, when new knowledge areas appear, new terms will not be present on dictionaries. The word alignment process of parallel corpora is very important to aid the extraction of specific multilingual terminology;
- automatic translation by studying human translations, developers can learn and infer new automatic translation algorithms. As translation resources, the sentence aligned corpora can be used to create translation memories to be used on MBMT (memory-based machine translation), and the full word aligned corpora can be used for EBMT (example-based machine translation);
- multilingual edition as an alternative to the automatic translation, the multilingual edition intends to generate different languages from a meta-

language: it is defined an artificial language L where all information possible is inserted, such that it is possible to generate diverse natural languages from it. This method can be effective when generating texts in a closed environment;

- product internationalization similar to automatic translation, but with a narrower focus;
- multilingual information retrieval systems that gather documents in different languages, where the query is written in any language (the original objective of Twente-aligner). This means that the query must be translated to all languages used on the database documents. As the translated query is not shown to the user, word-by-word translation based on translation probability can be used, with effective results;
- Contrastive linguistics: Parallel corpora are used to compare linguistic features and their frequencies in two languages subject to a contrastive analsis. They are also used to investigate similarities and differences between the source and the target language, making systematic, text-based contrastive studies at different levels of analysis possible. In this way, parallel corpora can provide new insights into the languages compared concerning language-specific, typological and cultural differences and similarities, and allow for quantitative methods of analysis.
- Translation studies: Closely related to the use of parallel corpora in contrastive linguistics is their application in translation studies. Parallel corpora may help translators to find translational equivalents between the source and the target language. They provide information on the frequency of words, specific uses of lexical items as well as collocational and syntactic patterns. This procedure may help translators to develop systematic translation strategies for words or phrases which have no direct equivalent in the target language. On this basis, sets of possible translations can be identified and the translator can choose a translation strategy according to the specific register, topic and genre. In recent times, parallel corpora have been increasingly used to develop resources for automatic translation systems.
- EFL-Classroom: Teachers are increasingly using parallel corpora in the classroom. In so doing they can determine the most frequent patterns of occurrence, enrich their personal knowledge of the language, design teaching

materials and provide authentic data in their teaching. Parallel corpora may also be helpful in the planning of teaching units and the identification of specific, potentially problematic, patterns of use and are thus useful tools for syllabus design.

- Moreover, parallel corpora can be used to identify translation difficulties and false friends. False friends are words or expressions of the target language that are similar in form to their counterpart in the source language but convey a different meaning. Even if words of the two languages have a similar meaning, they might belong to different registers or contexts, so that complete translational equivalence between source and target text is rare.
- Teachers are increasingly encouraging students to make use of parallel corpora themselves in order to become aware of nuances of usage and subtle differences in meaning.
- Lexicology: Parallel corpora are used more and more to design corpus-based (bilingual) dictionaries.

3.6.4 Corpora Creation in Indian Languages

The Central Institute of Indian Languages has corpora of around 3.5 million words for each major Indian language. The same will be enlarged to the extent of 25 million words in each language. Also, the existing corpora are raw corpora and it will be cleaned for use. Apart from 22 major Indian languages there are hundreds of minor and tribal languages that deserve attention from the researchers for their analysis and interpretation. Creation of corpora in these languages will help in comparing and contrasting structure and functioning of Indian languages. So, at least 100 minor languages corpora will be collected to a tune of around 3 to 5 million words in each language depending upon availability of text for the purpose.

Apart from these basic text corpora creations, an attempt are made to create domain specific corpora in the following areas:

1. Newspaper corpora
2. Child language corpus
3. Pathological speech/language data
4. Speech error Data

5. Historical/Inscriptional databases of Indian languages which is one of the most important to trace not only as the living documents of Indian History but also historical linguistics of Indian languages.
6. Grammars of comparative/descriptive/reference are needed to be considered as corpus of databases.
7. Morphological Analyzers and morphological generators.

3.6.4.1 POS tagged corpora

Part-of-speech (or POS) tagged corpora are collections of texts in which part of speech category for each word is marked. POS tagged corpora is developed in a bootstrapping manner. As a first step, manual tagging will be done on some amount of text. A POS tagger which uses learning techniques will be used to learn from the tagged data. After the training, the tool will automatically tag another set of the raw corpus. Automatically tagged corpus will then be manually validated which will be used as additional training data for enhancing the performance of the tool. This process will be repeated till the accuracy of the tool reaches a satisfactory level. With this approach, the initial man hours per 10,000 words will be more. Thereafter, the tagging process will speed up.

3.6.4.2 Chunked corpora

The chunking is done on the POS-tagged corpora. Here also the initial training set will be a complete manual effort. Thereafter, it will be a man-machine effort. Chunked corpora are a useful resource for various applications.

3.6.4.3 Semantically tagged corpora

The real challenge in any NLP and text information processing application is the task of disambiguating senses. In spite of long years of R & D in this area, fully automatic WSD with 100% accuracy has remained an elusive goal. One of the reasons for this shortcoming is understood to be the lack of appropriate and adequate lexical resources and tools. One such resource is the “semantically tagged corpora”.

In semantically tagged corpora, words in the text documents will be marked with their correct senses. For example apart from POS tagging, it is also necessary to tag the text with semantic tag to disambiguate homographic and polysemous words.

aTTai-1 ‘a living creature’

aTTai-2 ‘binding for a book’

maalai-1 ‘concerned with time’

maalai-2 <’that which is worn around neck’

The question that arises is “What should be the set of such tags and where should they come from?” Word Nets can be exploited for sense annotation. The IdowordNet consortia have decided to sense tag the corpus based on the wordNet ID number. This will be done manually in the training corpus which will used for testing corpus. By bootstrapping the size of the sense tagged corps will be increased.

3.6.4.4 Syntactic tree bank

Preparation of this resource requires higher level of linguistic expertise and needs more human effort. For preparing this corpora experts will manually tag the data for syntactic parsing. A tool can then automatically extract various tree structures for the tree bank. Since it requires more manual effort and also a higher degree of linguistic expertise, building of this resource will be a relatively slower process. The initial take-off time will also be more in this case.

Since, a crucial point related to this task is to arrive at a consensus regarding the tags, degree of fineness in analysis and the methodology to be followed. This calls for some discussions amongst the scholars from varying fields such as linguistics and computer science. It will be achieved through conduct of workshops and meetings. First some Sanskrit scholars, linguists and computer scientists will review the existing tagging scheme developed for Indian languages by IIIT, Hyderabad and define standards for all Indian languages (extendable to any

language). On this basis some experiments will be carried out on the selected Indian languages to test the applicability and quality of the defined standards. After testing, these actual tagging task will start.

3.6.4.5 Sources for Parallel corpora

A text available in multiple languages through translation constitutes parallel corpora. The National Book Trust, Sahitya Akademi are some of the official agencies who develop parallel texts in different languages through translation. Such Institutions have given permission to the Central Institute of Indian Languages to use their works for creation of electronic versions of the same as parallel corpora. The magazines, news paper houses that bring out translated versions of their output are another source to provide texts for parallel corpora. First wherever necessary the text have to be keyed in and then computer programmes have to be written for creating aligned texts, aligned sentences and aligned chunks.

3.6.4.6 Tools

The following tools are prepared for Indian languages under various consortia-projects.

1. Tools for Transfer Lexicon Grammar (including creation of interface for building Transfer Lexicon Grammar).

2. Spellchecker and corrector tools.

3. Tools for POS tagging. (Trainable tagging tool with an Interface for editing POS tagged corpora).

4. Tools for chunking (Rule-based language-independent chunkers).

5. Interface for chunking (Building an interface for editing and validating the chunked corpora).

6. Tools for syntactic tree bank, incl. interface for developing syntactic tree bank.

7. Tools for semantic tagging with basic resources are the Indian language WordNets showing a browser that has two windows - one showing the senses (i.e., synsets) from the WordNet appear in the other window, after which a manual selection of the sense can be done.

8. (Semi) automatic tagger based on statistical NLP (the preliminary version of which is ready in IITB).
9. Tools for text alignment, including Text alignment tool, Sentence alignment tool and Chunk alignment tool as well as an interface for aligning corpora.

3.6.5 Creating Multilingual Parallel Corpora for Indian Languages

Parallel corpora are of great importance in various natural language processing (NLP) and non-NLP tasks. Starting from a comparative and contrastive linguistic analysis for various linguistic features of the languages concerned to machine translation, there are various use for such a corpus in any given language pair.

India is nation with great linguistic diversity with over 452 individual languages listed by Ethnologue. Out of these, 22 languages are listed as 'scheduled' (also sometimes called 'national') languages comprising a total of 96.56% of the national population. Hindi is the largest spoken language across India (sharing above 41% of the national population) and also the official language of the Indian state (along with English).

Electronic content came rather late into Indian languages. The importance of corpus studies itself came into force with the prevalence of e-text. In such a scenario, the corpus study in Indian languages was negligible prior to this century. With the advent of common use of computers, the Indian languages also got some share and e-content gradually started growing in Indian languages. Though Unicode standards in Indian languages has helped grow the content, there is not enough content available that can be used to create parallel corpus in Indian languages.

There have been attempts to develop parallel corpora in Indian languages earlier as well. But none of such corpora have been developed from the scratch and

is mostly not publically available for the research community. Barring one exception of the EMILLE parallel corpus (Baker, P. et.al., 2004) of 200 thousand words in three languages in general domain, there is no other parallel corpus made in Indian languages. For the annotated parallel corpus, there are none available in Indian languages. To fill this gap, the Department of Information Technology (DIT), Govt. of India sanctioned a project run through a consortium involving 11 institutions across India (Jha, Girish Nath, 2010). It presents a summary of the work carried out under this project. This is an attempt to build a representative and comprehensive corpus of two domains in 12 major scheduled Indian languages.

These languages represent both the two major language families present in India, namely Indo-Aryan and Dravidian. Being the Associate Official Language (AOL) of India, English, a Germanic language, is also included.

The corpora creation has two principal tasks: creation of the raw parallel aligned text and POS annotation. The translation is done manually by especially trained native speakers of the language in their regions. Annotation is also done manually with no use of available automatic taggers.

For translation there are minimal guidelines with respect to format and structure of the target sentences. The source text is formatted to be one sentence per line and each sentence is given a unique identification (ID) number. The translated text in the target languages are also formatted accordingly i.e. they are one sentence per line and correspond to the sentence ID number of the source text. This ensures that we have the source and the target text aligned as we progress. We do not use any alignment tool for this purpose.

Creating the source text is equivalent to corpus creation. As the source text corpus is domain specific and has limitations with regard to the size each of these domains can grow, a careful selection of the text had to be followed. The two domains of health and tourism are not very prolific ones in Hindi. Most of the works done in these two domains are in English. Therefore finding original text in Hindi in these two domains has been a difficult task. The average of words per sentence (out

of a total of 25000 sentences per domain) comes out to be 16. Thus we get a corpus consisting of a total of about 400,000 words in each domain.

3.6.5.1 Creating the Source Text

While it is possible to collect the source text online, it is advisable that one should do this with extra caution when creating an ambitious corpus as presented here, particularly for less resourced languages like the Indian languages. Besides, most of the text over the internet would need editing and proofing. For the source text or the base corpus, they first tried selecting text online. But then they realized that most of the text that was available in Hindi over the internet was translated from English or other languages. Besides, our choice necessarily had to be very eclectic as they were specific about the domain and ensure that proper representation was given to the various sub-domains and genres within the domains. So, they went on to collect text from various other sources e.g. promotional materials published and distributed by government and/or private institutions/agencies. They also selected extracts from books, articles and stories from magazines and newspaper.

3.6.5.2 Domains of corpus

Initially the health-domain and tourism domain are taken for corpus collection for parallel-corpus collection among Indian languages.

3. 6.5.2.1 Health Domain

Health domain is divided into a total of 16 sub-domains. These sub-domains are made mainly to capture the different disciplines within the medical arena. No sub-domain is allotted to different genres of medical practice like allopath, ayurveda, acupressure, acupuncture etc. However, these are included in the corpus in a certain proportion with the total of the text. For example a disease, its description and symptoms are given only once as these are common in each of the medical practices. It is the diagnosis and treatment where the difference would be reflected.

The health domain has a total of 419420 words, with the total number of words per sentence being 16.77. The total number of unique words in this domain comes out to be 21446.

3.6.5.2.2 Tourism Domain

Tourism domain is divided into a total of 17 major sub-domains. These are further divided into categories as per the requirement. For example, pilgrimage is divided into two categories of Indian and extra-Indian, ecotourism is divided into wildlife, hill stations, desert and others. There are also sub-domains that do not have any categories like leisure tourism, medical tourism etc. Table 2 below gives a summary of the tourism corpus. The tourism corpus has a total of 396204 words with a per sentence word average of 15.8. Total number of unique words in the tourism corpus is 28542.

3.6.5.3 Data Storage, Maintenance and Dissemination

The Hindi source data collected manually with careful selection criteria in mind was mostly typed out by language editors. Out of the 25 thousand sentences in each of the domains only a meager 1500 sentences or 6% were taken from the internet. The whole of the corpus was first typed into spread sheets on normal PCs by the language editors of the source text. It was further validated by the present authors. Each sentence in the corpus has a unique ID which gets carried forward to each of the translated languages. Thus the alignment is done simultaneously as the translation in each of the languages progresses.

All the data collected and incorporated in the source text are stored with their metadata information which includes various information e.g. the source, number of words selected from the source, names of the authors/copyright holders and their sub-domain details. For the archiving purposes, all the source text is hyperlinked with a scanned image file of the source document from where the text was taken.

The source text is encoded in Unicode. All the translated texts in other languages are also in Unicode. As for the quality of the source or the translated text, we believe this to be the best possible.

For the translated text, usually only one translation faithful to the source text is expected. However, wherever possible, if two or more options are available for a sentence, the translators are encouraged to provide alternative translations as optional ones. The translated sentences are evaluated by external evaluators and the suggestions/corrections recommended by them have been incorporated in the target text. The whole of the corpus creation process has been supervised and the corpus principally has 0% 'noise' in terms of spelling mistakes, wrong character encodings, incorrect translations etc.

Govt. of India has started a data centre (http://tdil-dc.in) The ILCI corpora is in the process of being uploaded to this data centre and will be available for free download as per the Govt. of India guidelines.

3. 6.5.4 Parallel Corpus Creation

As noted above, the parallel corpora are created simultaneously, in each of the language pairs as the translation progresses. As the source text is created it is electronically sent to the other members of the consortium where the respective translators translate the source text in their respective target languages.

3. 6.5.5 POS Annotation

POS tagging is done on the translated corpus for each language. Although there are some POS taggers available for some of the Indian languages, their efficacy and standard input/output has been doubtful. Moreover, the POS tagset for Indian languages did not have a common standard till very recently when it got its first national standard in POS annotation through the efforts of BIS and ILCI.

3. 6.5.5.1 POS Tagset

There is no sizeable POS annotated corpus available in any of the Indian languages at present. As POS annotation is a part of this project, the tagset to be used for the corpora of these 12 languages became an issue. Several meetings were held under the aegis of BIS to come to a conclusion. Finally, a POS tagset was

agreed upon by the stake-holders. This tagset has come to be known as the BIS parts-of-speech annotation tagset. (No standard published reference can be given for this tagset as yet. We refer to the document circulated in the consortia meetings. This document was referred as “Linguistic Resource Standards: Standards for POS Tagsets for Indian Languages”, ver. 005, August, 2010.)

The BIS Tagset contains the features of the hierarchical tagset. However, it has tags for only first two tiers of linguistic information (POS and their subtypes) and excludes information from tier three onwards as these can be provided by morph analyzers and parsers. Morphological analyzers are available for some of the languages in the group and many more are in the process of being developed. For Hindi, morphological analyzers have been reported from various quarters e.g. (Goyal, V. & Singh Lehal, G. 2008; Bögel, T. et.al., 2007; etc).

3.6.5.5.1.1 Principles for Designing Linguistic Standards for Corpora Annotation

The BIS standard has set the following principles for designing linguistic standards for corpora annotation.

i. Generic Tag Sets

ii. Layered approach

- Layer I: Morphology
- Layer II: POS <morphosyntactic>
- Layer III: LWG
- Layer IV: Chunks
- Layer V: Syntactic Analysis
- Layer VI: Thematic roles/Predicate Argumentstructure
- Layer VII: Semantic properties of the lexical items
- Layers VIII, IX... Word sense, Pronoun referents (Anaphora), etc,

iii. Hierarchy within each layer

iv. Extensibility (including the language specific requirements and additional languages)

v. If a tag is redundant for a language, it should be deprecated

vi. ISO 639:35 Language code should be used <in metadata> 5 http://www.sil.org/iso639-3/default.asp

vii. Follow global guidelines such as EAGLES (Leech, G. & Wilson, A. 1999) where available.

viii. Standards should be mappable to/compatible with existing schemes to and from

ix. Standard is designed to handle wide range of applications and also should support all types of NLP Research efforts independent of a particular technology development approach

x. The scheme should be Annotator friendly.

3. 6.5.5.2 Super Set of POS Tags

Guided by the principles above, a super set of POS tags for Indian languages has been developed (Appendix I). Tagsets for different Indian languages have been drawn from this super tagset. As can be seen in Appendix I below, there are 11 top level categories. These are further classified into types and subtypes. There are a total of 45 tags in this set. If a language demands further sub-types, the principles above allow that. However, top level categories cannot be changed or new top level categories are not recommended to be added. No individual 4 language has used all of these categories. The tagsets for all the 12 languages have been drawn from this super tagset.

3. 6.5.5.3 Super Set of POS Tags for Indian Languages

3.6.5.5.4 Manual POS Annotation

The annotation is being done manually by the language experts/native linguists following the annotation guideline prepared for respective languages. There are some languages in the group that are morphologically agglutinating. For such languages direct annotation is not possible and morphological segmentation is required before POS annotation can begin. For such languages e.g. Tamil, Telugu and Malayalam, segmentation is recommended as a pre-processing task before the POS annotation. Additionally, a server-based, access-anywhere, annotation tool is

put in place where the annotators can annotate the text in their respective language over the internet. The tool can be accessed here: http://sanskrit.jnu.ac.in/ilciann/index.jsp

3.6.6 Creation of parallel Corpus for the SMT system

Here in this section the creation of parallel corpus for statistical machine translation (SMT) system will be briefly explained. SMT treats translation as a machine learning problem. This means that we apply a learning algorithm to a large body of previously translated text, known variously as a parallel corpus, parallel text, bitext, or multitext. The learner is then able translate previously unseen sentences. With an SMT toolkit and enough enough parallel text, we can build an MT system for a new language pair within a very short period of time.

Formally, our task is to take a sequence of tokens in the source language with vocabulary VF, and transform it into a sequence of tokens in the target language with vocabulary VE. We will assume that tokens are words and sequences are sentences. Agglutinative languages like Tamil may require special preprocessing. The most important consideration is that all data are preprocessed consistently, since statistical systems are sensitive to discrepancies. There is often no special treatment of morphological variants—for instance, the English words translate and translation are treated as unrelated, indivisible tokens. Therefore, it is possible for the size of the vocabularies VE and VF to reach into the tens or hundreds of thousands, or even millions in the case of morphologically complex languages such as Tamil.

Statistical machine translation is based on the idea that portions of any sentence can be found on other texts, specially, on parallel ones. We can say this is not the real truth, but happens for most of the cases. Relying on this idea, the statistical translation aims to divide a sentence on small chunks (three, four or more words) and search on parallel corpus for those sequence occurrence. Found them, the word sequence alignment algorithm can be used to determine the corresponding translations. Optimally the translation for those chunks (with overlapping words) can be composed together to form acceptable sentence translations. Of course the good translations observed on section 5.6 occurred because the word sequences appear in the corpus, and most cases, more than one time.

With this in mind, we have to develop a statistical translator prototype. For each sentence sα in the text we want to translate, we have to split it into its constituent words (or tokens, as punctuation is considered a word in this case): wα,1 . . . wα,n. Then, until there are no words to translate, we take a sequence of k words wα,i . . . wα,i+k−1 (normally 3 or 4) starting with i = 0 and try to find that sequence on the base corpus we are using for the translation. If the sequence is found, its aligned segment is added to the translation, and we restart the process with i = i + k. If not found, we take a smaller size segment (k = k − 1) and retry the alignment. This process is done until we find a segment to align (in the last case, when we find a word to align).

3.6.6.1 Corpus Collection

Corpus collection is a crucial issue in building an MT system based on statistical approach. Corpus collection should address the following issues:

- What parallel corpora look like.
- To view parallel corpora through the eyes of a computer.
- How parallel corpora are relevant to machine translation.
- How to build bilingual dictionaries from parallel corpora.
- How cognate information may be useful in machine translation.
- How to do word alignment, and how to employ the pigeonhole principle.
- About the chicken-and-egg nature of dictionaries (which enable word alignments) and word alignments (which enable dictionary building).

The following steps have been followed for the acquisition of a parallel corpus for the use in the present statistical machine translation system:

1. obtain the raw data (e.g., by crawling the web)
2. extract and map parallel chunks of text (document alignment)
3. break the text into sentences (sentence splitting)
4. prepare the corpus for SMT systems (normalisation, tokenisation)
5. map sentences in one language sentences in the other language (sentence alignment)

3.6.6.2 Compilation of parallel corpora

The texts of a corpus are chosen according to specific criteria which depend on the purpose for which it is created. In particular, compilers have to decide whether to include a static or dynamic collection of texts, and entire texts or text samples. Questions of authorship, size, topic, genre, medium and style have to be considered as well. In any case, a corpus is intended to comply with the following requirements: (i) it should contain authentic (naturally occurring) language data; (ii) it should be representative, i.e. it should contain data from different types of discourse.

3.6.6.3 Alignment of a parallel corpus

In order to use a parallel corpus properly it is necessary to align the source text and its translation(s). This means that one has to identify the pairs or sets of sentences, phrases and words in the original text and their correspondences in the other languages. Parallel text alignment is important because during the translation process sentences might be split, merged, deleted, inserted or reordered by the translator in order to create a natural translation in the target language. In order to compare the original text and its translation(s), it is necessary to (re-)establish the correspondences between the texts. In the process of alignment, anchor points such as proper names, numbers, quotation marks etc. are often used as points of orientation. The degree of correspondence between the texts of a parallel corpus varies depending on the text type. For example, a fictional text may allow the translator a greater freedom than a legal one.

The alignment at sentence and word levels makes parallel corpora both more interesting and more useful. As long as parallel corpora exist, sentence aligned parallel corpora is an issue which is solved by sentence aligners. Some of these tools are available as open-source software, while others have free licenses for non-commercial use, and produce reasonable results.

Usually, alignment tools perform the alignment at sentence and word levels. Texts are sequences of sentences. To sentence align two texts is to create relationships between related sentences. The same idea can be used for the word

alignment process: sentences are sequences of words. So, the word alignment process will add links between words from the original and the translated text. Word alignment can be viewed in two different ways:

– for each word, in a sentence, find the corresponding word in the translated sentence. This means that, for each occurrence of a word, it has a specific word linked to it.

– for each word from the source corpus, find a set of possible translations (and its probability) into the target corpus.

This leads to a Probabilistic Translation Dictionary (PTD), where for each different word of the corpus we have a set of possible translations and their respective probability of correctness.

The following issues to be kept in mind while aligning the corpus:

- About word alignment and dictionary building at a larger scale.
- About phrase-to-phrase alignment, the norm in real translation data.
- About unalignable function words.
- The importance of knowing the target language (versus source) in making fluent translations.
- The importance of short sentence pairs (where alignment possibilities are restricted) in helping disambiguate/align longer sentence pairs.
- About locality in word order shifts.
- How to guess the meanings/translations of unknown words.
- About how much uncertainty the machine faces in working with limited data.

3.6.6.3 Sentence Alignment

Sentence alignment is usually a hard problem, but in our case it is simplified by the fact that the texts are already available in paragraph aligned format. Each paragraph consists typically of only 2–5 sentences. If the number of paragraphs of a speaker utterance differs in the two languages, we discard this data for quality reasons. The alignment of sentences in the corpus is done with an implementation of the algorithm by Gale and Church [1993]. This algorithms tries to match sentences of similar length in sequence and merges sentences if necessary (e.g. two short

sentences in one language to one long sentence in the other language), based on the number of words in the sentence. Since there are so few sentences per paragraph, alignment quality is very high. There is considerable work on better sentence alignment algorithms. One obvious extension is to not only consider sentence length, but also potential word correspondences within sentence pairs. Work by Melamed [1999] is an example for such an approach. The sentence aligned data is stored in one file per day, so that lines with the same line number in a file pair are mappings of each other. The markup from the document aligned files is stripped out.

The alignment at sentence and word levels makes parallel corpora both more interesting and more useful. As long as parallel corpora exist, sentence aligned parallel corpora is an issue which is solved by sentence aligners. Some of these tools are available as open-source software, while others have free licenses for non-commercial use, and produce reasonable results. Regarding word level alignment, there are many interesting articles about the subject, referring many tools (Melamed, 2000; Hiemstra, 1998; Ahrenberg, Andersson, and Merkel, 2000). Unfortunately, most of them are not opensource nor freely available. Those that are available do not scale up to the size of corpora most researchers wish to align. With this in mind, word alignment is one area where there is still a direneed of research. Thus, this dissertation focuses upon the creation of better tools concerning word alignment.

For us, it is very important that the software used and developed follows the open-source philosophy. Without an open license, we cannot adapt the software to bigger applications, study the algorithms and implementations used or correct bugs. We can chose the open-source word aligner to help the bootstrap process for a parallel corpora package. Starting with a working software tool saves a lot of time, which can be applied to more interesting work, as there is no need to develop the application from scratch.

3.6.6.4 Word alignment

Parallel corpora are valuable resources on natural language processing and, in special, on the translation area. They can be used not only by translators, but also analyzed and processed by computers to learn and extract information about the

languages. Some processes related with the parallel corpora life cycle and the parallel corpora word alignment.

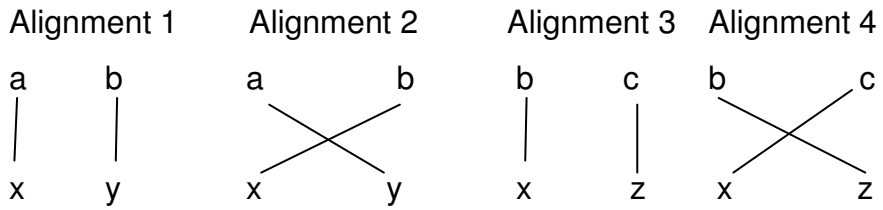
The necessity for a robust word aligner arrived with the TerminUM project which goal is to gather parallel corpora from different sources, align, analyze and use them to create bilingual resources like terminology or translation memories for machine translation.

The starting point was Twente-Aligner, an open-source word aligner developed by Djoerd Hiemstra. Its results were interesting but it worked only for small sized corpora. The work done began with the reengineering of Twente-Aligner, followed by the analysis of the alignment results and the development of several tools based on the extracted probabilistic dictionaries. The re-engineering process was based on formal methods: the algorithms and data structures were formalized, optimized and re-implemented. The timings and alignment results were analyzed.

The speed improvement derived from the re-engineering process and the scale-up derived of the alignment by chunks, permitted the alignment of bigger corpora. Bigger corpora makes dictionaries quality raise, and this makes new problems and new ideas possible.

The probabilistic dictionaries created by the alignment process were used in different tasks. A first pair of tools was developed to search the dictionaries and their relation to the corpora. The probabilistic dictionaries were used to calculate a measure of how two sentences are translations of each other. This naive measure was used to prototype tools for aligning word sequences, to extract multiword terminology from corpora, and a “by example” machine translation software.

Following could be the possible word alignments in the parallel corpus



3.7 Summary

Corpora is the term used on Linguistics, which corresponds to a (finite) collection of texts (in a specific language). Parallel corpora are valuable resources on natural language processing and, in special, on the translation area. They can be used not only by translators, but also analyzed and processed by computers to learn and extract information about the languages.

A collection of documents in more than one language is called a multilingual corpora. Multilingual corpora may be classified according to their properties. Parallel corpora is a collection of texts in different languages where one of them is the original text and the other are their translations. Comparable corpora are texts in different languages with the same main topic.

The first step in extracting useful information from bitexts is to find corresponding words and/or text segment boundaries in their two halves (bitext Maps). Bitexts are of little use, however, without an automatic method for matching corresponding text units in their two halves.

Although we can add morphological analysis, word lemmas, syntactic analysis and so on to parallel corpora, these properties are not specific to parallel corpora. The first step to enrich parallel corpora is to enhance the parallelism between units on both texts. This process is called “alignment”. Alignment can be done at different levels, from paragraphs, sentences, segments, words and characters.

.

Chapter 4

Parallel Structures of English and Tamil

4.0 Introduction

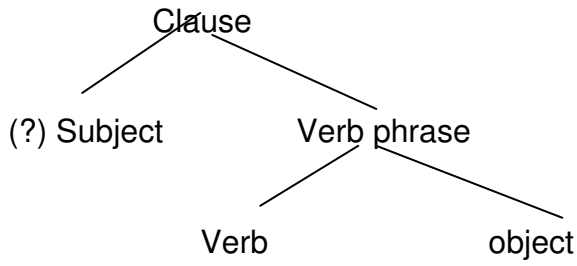
As the thesis is aimed at building an English-Tamil Machine translation system using statistical approach, there is no need to prepare a transfer grammar for English to Tamil transfer. The SMT system itself can find the parallel patterns for English-Tamil transfer and give the translated output. Even then while preparing the parallel corpus for the SMT system it is better understand the parallel structure between English and Tamil as it may help in preparing parallel corpus from simple to complex ladder. While building SMT system for English-Tamil transfer, we may give the system parallel corpus for training the system. In that context we have to give the

parallel corpus in phase wise manner of simple to complex. So the present chapter is aimed at finding the parallel structure between English-Tamil so that the SMT system can be trained properly. Kamakshi (Kamakshi and Rajendan 2004) has discussed in detail about the parallel structure of English and Tamil while she was building a transfer grammar for English and Tamil transfer using transfer approach of MT. Her data is made use of here to understand the parallel structure of English and Tamil.

The parallel structrures of English and Tamil are extracted from the parallel corpus created for Englihs-Tamil SMTsystem.

4.1. Parallel sentential structures of English and Tamil

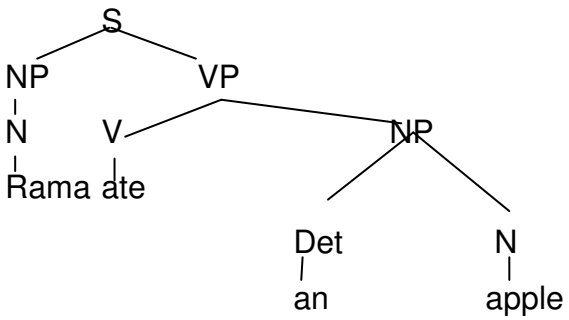
English relies on word order as a means of expressing grammatical relationships within constructions. In Tamil, word order is more flexible, as grammatical relations are signaled by inflections. In generative linguistics, English with fixed word order is called configurational language and Tamil with fairly free-word order is called non-configurational language. The core of the configurationality issue is about the question of special grammatical relation of subject and a different one of object, whatever these relations correspond to different positions in the hierarchy of the sentence. In Tamil, there is little or no evidence for a hierarchy as given below, but very often Tamil differentiates subjects and objects in crucial ways.



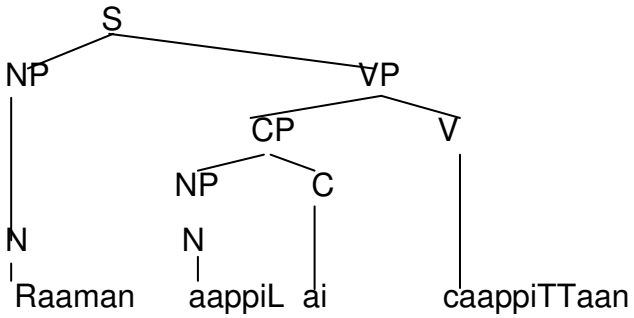
It has been taken for granted that in English there is a syntactic VP node. It is generally believed that Tamil lacks VP constituency. So generally Tamil sentences are given a flat structure without VP being at a different hierarchical level.

Subathra Ramachandran (1975) strongly argues that Tamil is a configurational language possessing a VP node. Even if it is true, we cannot compromise on the fact that Tamil is different from English as English is an SVO language where as Tamil is an SOV language in which 'S' and 'O' can be shuffled. Tamil is not strictly a configurational language. The object is decided by position in English whereas in Tamil by case markers.

English: Rama ate an apple



Tamil: raaman aappiLaic *caappiTTaan*.



Many interesting points will be revealed for the purpose of transferring English language structure into Tamil, if we look at the correlating features of the two languages from the point of view of their typological characteristics as SOV and SVO languages respectively.

1. Syntactically, English and Tamil are perhaps most saliently different in the basic word order of verb, subject and object in simple declarative clauses. English is an SVO language, meaning that the verb tends to come between the subject and object and Tamil is an SOV language, meaning that the verb tends to come at the end of basic clauses. So the two languages differ in their ordering of certain functional units. For example, English being an SVO language has prepositions, whereas Tamil being SOV language has postpositions.

2. The affirmative sentence in English which are in SVO order becomes aux + SVO to form interrogative sentences which is a discontinuous order. In Tamil, the interrogation does not change the word order.

3. English is a highly consistent SVO language. The government constructions observe SVO patterns, as do the nominal modifying constructions – with the exception of descriptive and limiting adjectives in an archaic order. As a consistent language, English exemplifies characteristic features of SVO languages, such as the many patterns that have been developed in the verbal modifying constructions, the wide use of substitutes, and the grammatical processes used to highlight elements of sentences. The verbal patterns make heavy use of auxiliaries, which are also involved as substitutes and in interrogative and negative constructions, differentiating English in this way from (S) OV languages like Tamil. The grammatical process involves function words, again in distinctive constructions like clefting.

Tamil is a typical (S) OV language in which the verb occurs at the final position of a sentence. Word order in the sentence is relatively free, as long as the sentence ends with a main verb. For example, the sentence *Kannan introduced Uma to Raja* in Tamil can have the following word- order- variants.

1. kaNNan umaavai raajavukku aRimukappa Tuttinaan.
   1 2 3
2. kaNNan raajaavukku umaavai aRimukappa Tuttinaan.
   1 3 2
3. umaavai raajaavukku kaNNan aRimukappa Tuttinaan.
   2 3 1
4. raajaavukku umaavai kaNNan aRimukappa Tuttinaan.
   3 2 1
5. raajaavukku kaNNan umaavai aRimkappa Tuttinaan.
   3 1 2
6. umaavai kaNNan raajavukku aRimukappa Tuttinaan.
   2 1 3

ai and *kku* are accusative and dative case markers and nominative is unmarked in Tamil. The above sentences are identical in logical content, but are different in

discourse presupposition in a very subtle way. Ordinarily, constituents that represent older information precede those that represent newer information. The subject – initial sentence pattern is the most common among the various word order patterns. In declarative sentence with nominal subject and object, the dominant order is almost always one in which the subject precedes the object.

4. Simple, unmarked clauses in English agree with the SVO pattern, and require representations for the three constituents: subject, verb, and object. Neither the subject nor the verb nor the object of a transitive verb may be omitted.

Uma folded her hands.
*Her hands Uma folded.

5. English does not permit any order other than the above in unmarked sentences occurring as single utterances. This constraints applies also in subordination, as in the following sentences:

Kannan shouted while Uma folded her hands.

6. In English the verbal qualifiers must precede verbs. This position conflicts with the optimum position for subjects. To express negation, for example, the negative element might be prefixed to the verb.

Uma does not fold her hand
*Uma folded not her hand
In Tamil the negative element follows the verb
*Umaa tan kaikaLai maTikkavillai Uma did not folded her hands'
*Umaa tankaikaLai illaimaTittaaL

7. Government operates strongly in English, both in predicates and in other government constructions.

Her hands are folded.
Two of her hands are folded.

This is case with Tamil too.

avaL kaikaL maTikkappaTTana ‘Her hands are folded’

avaL iru kaikaLum maTikkappaTTana ‘Two of her hands are folded’

8. In comparison of inequality the adjective precedes the standard.

Uma is more beautiful than Usha.

9. In titles, the name follows, functioning like a standard for the ‘variable’ title.

Queen Uma.

Tamil allows both the possibilities.

raaNi umaa ‘Queen Uma’

umaa raaNi ‘Queen Uma’

In personal names the surname follows as standard to the given name.

John, F Kennedy.

10. In numerals in the teens, the form of ten follows, as ten follows, as in the other constructions of this kind furnishing a sturdier for the simple numerals from three to mine ex in Tamil pattern is tent numeral.

Thirteen, fourteen, fifteen, sixteen, seventeen, eighteen, nineteen.

patinonRu ‘eleven’, panniraNTu ‘twelve’, patimuunRu ‘thirteen’, patinaanku ‘fourteen’, patinaindtu ‘fifteen’, patinaaRu ‘sixteen’, patindeezhu ‘seventeen’, patineTTu ‘eighteen’, pattonpatu ‘nineteen’

11. English has been characterized by functional syntacticians as a language in which the initial segment, or theme, often using old material, sets the scene for the new material, or rhyme.

Uma folded her hands.

The subject Uma is one of the important elements of the preceding discourse, while the predicate folded her hands introduces a new action. SVO order provides a convenient basis for such organization of sentences. The same can be said for Tamil too.

umaa tan kaikaLai maTittaaL ‘Uma folded her hands’

12. For the basic sentential structures identified for English, the corresponding Tamil structures are given.

English	Tamil
SVA Arul is in the reception hall	S AV aruL varaveeRpaRaiyil irukkiRaan
SVC Arul is clever	SCV aruL puticaali aavaan
SVO Arul threw the ball	SOV aruL Pandtai eRindtaan
SVOA Arul kept the ball on the table	S O A V aruL pandtai meecai meel vaittan
S V O C Arul has proved her wrong	S enRu-clause V aruL avaL tavaRu enRu ndiruupittaan 'Arul has proved that is wrong'
S V O I O Arul taugh her music	S I O O V aruL avaLukku icai kaRpitaan
SV The baby cried.	SV Kuzhandtai azhutatu

(Here in this context A = Adjunct, C = Complement, I O = indirect Object, O = Object S = Subject, V=Verb)

In both English and Tamil simple, compound and complex sentences have been identified traditionally

Sentence	English	Tamil
1.Simple sentence	He goes to market	avan *maarkeTTukkup pookiRaan*
2. Compound sentence	He went to market and bought some vegetables	avan *maarke TTukkup poonaan kaaykaRikaLai vaangkinaan*
3. Complex sentence	He is going to market to buy vegetable	Avan kaaykaRivaangka candtaikkup pookiRaan

Though the distinction of sentences into simple, compound, complex is traditional, it is crucial from the point of view of translation. We can expect a simple sentence in the source language having an equivalent simple sentence in the target language. For example, for a simple sentence in English, we can expect a simple sentence as its translation equivalent in Tamil. Similarly for a compound sentence in English we can expect a compound sentence as its translation equivalent in Tamil.

Also we can expect a complex sentence as a translation equivalent in Tamil for a complex sentence in English.

English	Tamil
He went to market NP V to-NP	avan candtaikkup poonaan NP NP-ukku V
He went to market and bought vegetable NP V to – NP V NP	avan candtaikkup poonaam maRRum kaaykaRikaL vaangkinaan. NP Np-ukku V maRRum NP V
He went to market to buy vegetable. NP V to-NP V NP	avan kaaykaRikaL vaangkad Candtaikkup poonaan NP V-INF NP-ukku V

But this idealization of getting translation equivalent as mentioned above may not be true always. One can expect a complex sentence for a simple sentence or vice versa.

English	Tamil
Arul has proved her wrong NP V NP Adj	arul avaL tavaRu enRu Ndiruupittaan 'AruL has proved that she wrong' NP [NP NP] S enRu V

This complexity should be kept in mind while looking for translation equivalents of English sentences in Tamil.

Traditionally in both English and Tamil the following types of sentences are identified.

Sentence type	English Sentence	Corresponding Tamil sentence
Affirmative or assertive sentences	He went to market	*avan candtaikkup poonaan*
Question or interrogative sentences	Is he going to market?	*avan candtaikkup pookiRaana?*
Negative sentences	He is not going to market	*Avan candtaikkup Pookavillai*
Imperative or command Sentences	Go to Market	*Candtaikkup poo*
Exclamatory sentences	How beautiful the building is!	*aa!evvaLavu azhakaana kaTTiTam itu1*

This functional distinction of sentences is also crucial to our venture in finding translation equivalents in Tamil for English sentences. It should also be noted that the word order plays a crucial part in converting affirmative sentences into interrogative sentences in English. In Tamil word order does not play a crucial role while transforming an affirmative into an interrogative sentences, it makes use of clitics.

4.1.1 Parallels of affirmative sentences

English has an explicit link verb ('be' verb) to equate the subject NP with the complement, NP, Adj, and Adv. Explicit link verb is lacking in Tamil. But there are contexts in which we make use of 'be' verb, which can be equated with English 'be' verb. The 'be' verb *iru* can complement an NP through an adverbial particle *aaka*.

avaL oru maruttuvaraaka irukkiRaaL
'She is a doctor'
avaL azhakaaka irukkiRaaL
'She is beautiful'
avaL cennaiyil irukkiRaaL
'She is in Chennai'

Adjective in Tamil cannot occupy the predicate position as in English. (In English adjective is supported by the 'be verb). In Tamil *aaka* helps in the formation of an adjective when followed by the verb *iru*. The following table will depict the mechanism of transfer of equative sentences in English into Tamil.

Structure of English equative sentences	**The corresponding structure of Tamil equative sentences**
NP + 'Be' verb + NP Kala is a girl	NP + NP Kalaa oru ciRumi
NP + 'Be' verb + NP Kamala is a doctor	NP + NP-aaka + iru-T-PNG Kamalaa maruttuvaraaka irukkiRaaL
NP + Be verb + Adj. Kamala is beautiful	NP + NP-aaka + iru-T-PNG Kamala azhakaaka irukkiRaaL NP + NP-aana-PN Kamala azhakaanavaL

NP + Be verb + Adv. Kamala is there	NP + Adv. + iru-T-PNG kamala angkee irukkiRaaL
NP + Become + NP Kamala became a teacher	NP + NP + aaku-T-PNG Kamala aaciriyar aanaaL

In Tamil, the equative sentences of NP + NP type are used in the present context. If the equation is made in the future and past contexts, Tamil needs the help of the 'be' verb *iru*, which can be inflected for past and future.

Kamalaa oru maruttuvar
"Kamala is a doctor'
Kamalaa *oru maruttuvaraaka irundtaaL*
'Kamala was a doctor'
Kamalaa *oru maruttuvaraaka iruppaaL*
Kamalaa will be/may be a doctor.

4.1.2. Parallels in interrogative sentences

An auxiliary is proposed before the subject to express interrogation in English.

Did he come yesterday?
Do cats eat bats?

Such questions require an answer of either yes or no, and as a result they are often labeled yes-or-no- questions.

In accordance with the general principle, the interrogative marker should stand close to the sentence boundary, whether initially in VO languages or finally in OV languages. English makes use of a special set of words, which may combine with the interrogative with a substitute for the subject, the so-called wh-words. For yes-or-no questions it has lead to the introduction of auxiliaries. Among the auxiliaries *do* is the most remarkable in having today only a grammatical function, whether as interrogative marker or as a device for the indication of negation or emphasis. Other auxiliaries combine uses as grammatical markers with expression of modality, aspect and tense.

The second set of questions in languages is characterized by a question word. These often referred as wh-question words after wh-segment in many English interrogative words, because they include an 'unknown quantity'. Initial position of the interrogative element accords with the expectation of this order for the theme as well as with the general ordering principle.

What's the French word for cuckoo?
What right has you to call me uncle?
How is bread made?

English, as an SVO language, permits only one wh-word before finite verb, whether this is a noun, an adjective, or an adverb.

Besides wh-questions and yes-or-no questions, English includes devices indicating presupposition in yes-or-no question. One such device is tag question, consisting of a positive auxiliary when a negative answer is presupposed, and a negative auxiliary for a presupposed positive answer. The auxiliary corresponds in form to that of the principal verb, as in the following idiosyncratic statement.

I speak English, don't I?

Interrogative expressions are then closely related to expressions for sentence negation, though negation may be used for syntactic rather than pragmatic purpose.

The three types of interrogation found in English have to be correlated with that of Tamil for the purpose of developing MTA.

4.1.2.1 Parallels in yes-no questions

Contrasting characters pertinent to the transfer of yes-no questions in English into Tamil needs close scrutiny. As we have already noted, the yes-no questions can be sub divided into three types in English:

1. Those with 'be' verb

2. Those with ‘modal’ auxiliary
3. Those with ‘do’ verb

The table below correlates the question with ‘be’ verb in English with Tamil.

English pattern	The corresponding Tamil pattern	Comments
1.Question with be verb 1.1.Be + NP + NP Is she a teacher? 1.2.Be n’t + NP + NP Isn’t she a teacher? 1.3.Be + NP + ADj Is she beautiful?	1.1.NP + NP-aa avaL oru aaciriyaraa? 1.2.NP + NP + allav/illaiy-aa? avaL aaciriyar allav/ ilaiy-aa1 .3.NP + NP-aana-PNG-aa avaL azhakaanavaLaa	In the case of negative sentence with *not*, the short form n’t will be placed after the ‘*be’* verb. The movement of English ‘be’ verb to the initial position, is matched by adding of clitic *aa* in Tamil.
2.Question with modal auxiliary 2.1.Modal + NP + Main verb + (NP) Can he be a doctor? 2.2. Modal + NP+V Can I sing? Should I write?	*NP-aal + NP-aaka + iru + Modal-aa* *avanaal maruttuvaraaka irukka muTiyumaa?* *NP-aal + V-INF + Modal-aa* *ennaal paaTa muTiyumaa?* *ndaan ezhuta veeNTumaa?*	
3.Question with do verb 3.1.Do + Tense + NP + V + (NP) Did write the story? 3.2. Don’t + Tense + NP + V + (NP) Didn’t Rani write the story?	3.1.NP + (NP) + V-T-PNG-aa raaNi katai ezhutinaaLaa? 3.2.NP + (NP) + V-INF iilaiy-aa raaNi katai ezhutavillaiyaa?	The movement of English *do* to the initial position (or do insertion) is matched by adding of clitic *aa* in Tamil

Interestingly to trigger all the three types of interrogation in English, Tamil makes use of the clitic *aa* with the relevant units. It can be summarized that for the question type in English where the ‘be’ verbs such as *is*, *am, was, are, were, will be, shall* be are proposed to the subject to frame questions, we can expect two types of equivalents in Tamil.

The interrogative structure of type ‘be’ verb + NP+NP in English will be matched by ‘NP + NP-aa’ in Tamil.

Is she a girl?
avaL oru ciRumiyaa?

For the interrogative structure of the type “be’ verb + NP + Adj’ in English, Tamil makes use of NP + [NP-aana] Adj-PN-aa. Note that in Tamil adjective is pronominalized when it is used as a predicate.

Is she beautiful?
avaL azakaanavaLaa?
Is he poor?
avan eezhaiyaa?

The yes-no question of the type “be’ verb + NP + adv’ in English is matched by ‘NP + Adv + iru-T-PNG-aa’ in Tamil.

Is he here?
avan inkee irukkiRaanaa?

4.1.2.2 Parallels of wh-questions

The ‘wh’ questions of Tamil are similar to ‘wh’ questions in English. As in English, interrogative pronouns, adverbs, etc in Tamil introduce them. Since the word order is flexible the question word can be introduced anywhere in the sentence. As we have noticed already, English question sentences are formed by the movement of the operator verb followed by the movement of interrogative words (noun, adverb, adjective). Tamil does not have operators to be fronted although it has counterparts for ‘wh’ words in English. These counterparts do not move to the front of the clauses. As we noted already ‘yes/no’ questions in Tamil differ from their declarative counterparts by suffixing a clitic to the concerned elements to be questioned.

The following table shows the correspondence between interrogative words in English and Tamil.

Interrogative words in English	Interrogative words in Tamil
Who	yaar, evan, evaL, evar
Which, What	etu, evai, enna
How much	evvaLavu
How many	ettanai
How	eppaTi, evvaaRu
Where	engku
When	eppootu, eppozhutu
At what (time)	endndeeram, eppootu
On which (day)	e(endaaL), enRu
At what (place)	engkee
In which (town)	e(endakaril)
To which (country)	e(endaaTTil)
By whom	yaaraal
With whom	yaaruTan
With which (friends)	enta (ndaNparkaLuTan)
Whose (house)	yaaruTaiya (viiTu)
Why	een, etaRku

The following table shows the correspondence between interrogation in English and Tamil.

Interrogation in English	Interrogation in Tamil
In English interrogation is framed by shifting the auxiliary verbs to the initial position of the construction or by introducing the interrogative words or by adding super segmental morphemes to any statement. Interrogation framed by auxiliary for yes-or-no type. In English the auxiliary verbs and the interrogative words occur in the initial position. Tag questions are framed by auxiliary movement. In interrogative sentences,	Interrogation is framed by the addition of interrogative particles or by the introduction of interrogative words or by adding super segmental morphemes to any statement. Interrogation is framed by interrogative particles for yes-or-no type In Tamil the interrogative particles occur in the final position of any word in the construction, the interrogative words occur in the initial, medial or final position of the construction. Tag questions are framed by suffixing the interrogative clitic the concerned units. In Tamil too, the interrogative words occur in isolation in interrogative sentences.

interrogative words occur in isolation.	

4.1.3 Parallels in negative sentences

Under this title the transfer of negation in equvative sentences and non-equvative sentences and transfer of negative pronouns and determiners are dealt here.

4.1.3.1 Parallels in negation in equvative sentences

In the following table, negation in equvative sentences in English is correlated with that of Tamil.

English	Tamil
NP + BE-V +not + NP She is not a school teacher	NP + NP + *illai* avaL oru paLLi aaciriyar illai
There + BE-V + no + NP There is no God	NP + *illai* kaTavuL illai
NP + BE-V + not + PP	NP-LOC + *illai* avan cennaiyil illai
He is not in Chennai There + BE-V + no + NP There are no girls	*angkee + NP + illai* angkee maaNavikaL yaarum illai
It + is + not + ADJ-to-clause It is not easy to go there.	S-atu + adjectival noun + *alla / illai* angkee poovatu eLitallal/eLitu illai

4.1.3.2 Parallels in negation in non-equative sentence types

In the following table, negation in non-equvative sentences in English is correlated with that of Tamil.

English Negative forms	Corresponding Tamil forms
Did not +MV1 He did not go	MV-INF + *illai* avan pookavillai
Does not + MV1 He does not go	MV-atu + *illai* Avan poovatillai
Cannot + MV1 I cannot go	MV-INF *+ iyalaatu/muTiyaatu* ennaal pooka iyalaatu / muTiyaatu
Could + MV1 I could not go	MV-INF + *iyalavillai/muTiyavillai* *ennaal pooka iyalavillai / muTiyavillai*

Shall not + MV1 I shall not go	MV-INF + *maaTTu* + PNG/muTiyaatu ndaan pooka maaTTeen
Should not + MV I should not go	MV-INF + kuuTaatu ndaan pooka kuuTaatu
Will not + MV 1 I will rot go	MV INF + *maaTTu* + PNG ndaa Pooka maaTTeen
Would not + MV_1 I would not go	MV INF + *maaTTu* + PNG ndaan Pooka maaTTeen
Will not + MV1 It will not go.	MV-*atu* atu pookaatu
Must not + MV I must not go	MV-INF + kuuTaatu ndaan pooka kuuTaatu

4.1.3.3 Parallels in negative pronouns and determiners

English makes use of negative pronouns such as *none, nothing, neither, nobody, none* and negative determiner *no.* Tamil does not have negative pronouns; instead it makes use of *um*-suffixed interrogative pronouns such as *yaarum, etuvum, evarum* that do not posses any negative feature; the negation is expressed by adding *illai* to the verb.

Negative pronouns in English	**Corresponding pronouns in Tamil**
1. Person = no one, nobody No one come Nobody come	*Yaarum* *yaarum varavillai.*
2. Non-personm = nothing I ate nothing	*onRum* *ndaan onRum caappiTavillai*
3. None None of the students is good. None of the students here arrived	Oruvarum/yaarum *maaNavarkaLil oruvarum nallavarillai* *maaNavarkaLil yaarum varavillai*.

Tamil does not have a negative word equivalent to addition adverbial negative form *neither… nor*. Tamil makes use of negative verb form such as *illai, muTi, kuuTaatu,* V-aatu. The noun phrases or pronouns will be added with the disjunctive clitic *oo.* Similarly Tamil does not have the equivalent of *never* which is a negative adverb. Tamil makes use of Gerundive form of the type V-T/N-atu.

English	**Tamil**
Neither … nor	NP-*oo* NP-oo MV-INF _ *illai*

Neither Usha nor Uma came today	uSaavoo umaavoo inRu varavillai
Never Never I met him	V-T-RP-*atu+illai* ndaan avanai candtittatillai Note : The adverb *orupootum* which can be equated with English *never* also need the support of the negative verb *illai. orupootum* can be compensated by the emphatic clitic *ee* added to the VN. ndaan avanai orupootum candittatillai. ndaan avanai cantittatee illai.

The following table sums up the correlative features of English and Tamil for the sake of computation.

Negation in English	**Negation in Tamil**
Negation is effected by the addition of segmental word or suprasegmental morphemes to the affirmative construction.	Negation in Tamil. In Tamil too, negation is effected by the same method.
Forms like *no, not, nothing, nobody, none, not only, rarely, scarcely, seldom, few, little,* etc. are the negative words which are used to bring in negation.	The negative roots *al, il, maaTTu* are the negative words which bring out negation in Tamil.
The negative words do not show concord with the subject of the negative construction.	Except the negative verbs *alla and illai,* all other inflected negative words (ex. *maaTTu*) show concord with the subject of the negative construction.
Monomorphemic negative words in pure negative constructions occur with auxiliary verbs and others occur without any auxiliary verbs.	The negative words follow the nouns, a few adverbs or the infinitives.
The negative words in pure negative constructions occur in the initial position of the construction	The negative words in near negative constructions occur in the medial position.
A construction with a main verb can be negated only after introducing an auxiliary verb.	
The monomorphemic negative word in pure negative constructions is moved with the auxiliaries to form negative interrogative constructions whereas the negative words in near negative constructions are not moved with the auxiliaries.	

4.1.4 Parallels in imperative sentence

Imperative sentences are associated mostly with the second person commands, instructions and requisitions. Both Tamil and English make use of verb root to express impetration. As English does not have overtly marked separate terms for singular and plural second persons and the verbs do not inflect for subject, it does not have separate verbal form for singular imperative and Plural imperative. As Tamil has two or three distinct second person pronouns, one expressing singular *(ndii),* another expressing plural *(ndiir)* and honorific *(ndingkaL)*, it reflects this distinction in the imperative forms of verbs too. So, for English *you*, depending upon the context, Tamil may have at least two forms, one is verbal root and another is 'verb root + ungkaL'. The following table correlates different ways of bringing imperative sense in English and Tamil:

Imperative type	**English pattern**	**Corresponding Tamil pattern**
Imperative positive	Non honorific: V1 (i.e. verb root) Go Emphasis: Do + V1 Do go Honorific Please + V1 Please go Be + adjectival Compliment Be quiet Be serious Be seated	Imperative singular: Verb root Poo V-Past participle + *viTu /koL* pooyviTu, pooykkoL Imperative plural / Honorific Verb root + ungkaL poongkaL N-aaka + iru Amaitiyaay iru Kavanamaaka iru uTkaarungkaL
Imperative negative	Don't + V1 Don't go Don't sit	Verb root –aat- ee pookaatee uTkaaraatee
Let command	Let + object pronoun (1st & 3rd)/ Proper Noun + V1 Let him go Let me go Let Rama go	Sub (1st & 3rd person) /Proper noun _ V-INF – *aTTum* avan pookaTTum ndaan pookaTTum raaman pookaTTum
Positive *should* command	2nd person + should + V1 You should go	Sub (all three persons) + V-INF + veeNTum ndii pooka veeNTum

Negative *should* command	Sub (all three persons) + should not + V1 I should not go	Sub (all three persons) + V-INF + *kuuTaatu* ndaan pookak kuuTaatu
Positive must command	2nd person All person + must + V1 You must go	Sub (all person) + V-INF
Negative must command	Sub (all three persons) + must not + V1 You must not go	Sub (all three persons) + V-INF + kuuTaatu ndii/ndiingkaL Pookakkuu taatu
Positive has to/ have to command	Sub (all three persons) + has to/ have to + V1 + You have to go He has to go	Sub (all three persons) + V-INF ndaan pooka veeNTi irukkum avan pooka veeNTi *irukkum*
Negative have to command	Sub (all three persons) + don't/doesn't have to + V1 I don't have to go He doesn't have to go	Sub (all three persons) + V-INF + veeNTiyirukkaaatu ndaan pookaveeNTirukkaatu avan pooka veeNTirukkaatu
Positive need command	Sub (all three perons) + need + infinitive He need to go	Sub (all three persons) + V-INF + veeNTum ndii pooka veeNTum
Negative need command	Sub (all three persons) + need not + V1 He need not go	Sub (all three persons) + V-INF + veeNTaam ndii pooka veeNTaam

4.2. Parallel Clause Structures of English and Tamil

Three important types of subordinate clauses and the four nonfinite sub classes that can be structurally identified for English are correlated with that that of Tamil.

Type	**English**	**Tamil**
Finite clause	As + S+S As she is ill, she cannot attend her class.	*atu* – clause – *aal* + S uTalndalam illatataal avaLaal vakuppiRkucc cella iyalaatu
Non-finite clause	1. Infinitive clause with *to* It is better for you to do that work. 2. Infinitive clause without *to* All he did was open the door. 3. ing – partcipal clause	1. *atu-clause* + S ndii anda veelaiyaic ceyvatu ndallatu. 2. atu – clause + S avan ceytatu ellaam katavaitt

	Having arrived late, he was disappointed. 4. ed-participal clause They hurried home, the work is completed.	tiRandtatu taan. 3. Verbal participle clause + S taamatamaaka vandtu avan eemaandtaan. 4. S [… V-finite + S [… V-finite] S avarkaL viraivaakac cenRaarkaL. Veelai muTindtu viTTatu.
Verbless clause	[Although + ….]Verbless clause + S Although very cheerful, mary has many problems.	aalum-clause + S mika makizhcciyaaka irundtaalum, meerikku pala cikkalkaL uNTu.

Dependent clause may function as subject, object, complement or adverbial.

Dependent clause and they in English	**Equivalent clauses and their functions in Tamil**
[That …] NP + VP Subject That he is lazy is a fact.	S + *enpatu* + NP Subject avan coompeeRi enpatu uNmai
NP V [that ….] NP Direct object I know that he likes you.	S + *enRu* + S Direct object avan unnai virumpukiRaan enRu enakkut teriyum.
NP BE [that …] NP Subject complement The point is that he is your friend.	S + *enpatu* + NP avan unnuTaiya ndaNpan enpatu kuRippu.
NP + V interrogative clause {Io} + NP {o} Indirect object I gave whoever it was a cup of tea.	… V – aalum clause + S yaaraaka irundtaalum ndaan oru koopai teendiir koTutteen.
S + participle clause He found her excited with joy.	avaL makizhcciyaaka irukka avan kaNTaan
Adverbial clause + S Adverbial When we meet, I shall explain it.	V-um + pootu clause + S ndaam candtikkum pootu ndaan atai paRRi viLakkuveen

Subordinate clauses in English and Tamil can be correlated according to semantic criteria – whether their perspective on the content of the superordinate clause, as indicated by their subordinator, is one of time, location, manner, reason, etc.

Subordinate clauses in English	**Perspectives**	**Subordinate Clause in Tamil**
S + After + S	TIME	… V + Past + adjectival

ex. I questioned them after Usha met them.		participle + *pinnar*/ *piRaku* + S uSaa avarkaLai candtitta pimmar/piRaku ndaan avarkaLiTam keeLvi keeTeen
S + before + S I saw Ramu before he died	TIME	V + *um* + *mun* Raamu caakum mun ndaan avanaip paartteen
S + since + S I am in contact with them since I last met them.	TIME	… V-*atu* clause +*il+irundtu* + S ndaan avarkaLaik kaTaiciyaakac candittatilirundtu avarkaLiTam toTarpu vaittirukkiReen
S + until + S I will be staying in Chennai until we meet again	TIME	… V-*atu* clause + varai + S ndaam miNTum candtippatu varai ndaan cennaiyil iruppeen
S + when + S I hated you when you were in love with him.	TIME	Adjectival clause + pootu + S Ndii avaLiTam kaatal koNTirukkum pootu ndaan unnai veRutteen
S + while + S I was watching them while they were eating.	TIME	Adjectival clause + pootu + S avarkaL caappitukiRa *pootu* ndaan avarkaLaip paarttuk koNTirundteen.
S + where + S I don't know where Sujatha went.	LOCATION	…*engku* + V-finite + *enRu* + S cujaataa engku cenRaaL *enRu* enakkat teriyaatu
Since + S Since you didn't pay Money, we will not Give you books	REASON	*aal*-clause of ndii paNam kaTTaatataal ndaangkaL unakkup puttakam tara maaTTom
S + as if + S HC acted as if usha was Sick	MANNER	*atu* – clause + Pool + S Usaa uTal ndalamillaatatie Pool avan ndaTittaan
S + as though + S He acted as though Kannan hadn't seen him	MANNER	Atu-clause + *pool* + S kaNNan tannaip paarkkaatatu pool avan nda Tittaan.
S+need + S I need a bag so that I can take it safely.	PURPOSE	*ataRkaaka*-clause + S *atai* paatukaappaaka eTuttuc celvataRkaaka enakku oru pai veeNTum
S + in order that + S I encouraged Rajan in order that he would confer.	PURPOSE	…*um*-clause + *paTi*+ S raajanaik kalandtaaloocikkum ndaan uRcaakap paTuttineen
S + so as to + V1… He went through a	PURPOSE	*ataRkaaka* clause + S avan kaavalarait tavirppataRkaaka

small lane so as to avoid the police		oru kuRukiya paatai vazhiyaakac cenRaan
S +in order to+V1… He went through a small lane in order to avoid the police	PURPOSE	*ataRkaaka* clause + S avan kaavalarait tavirppataRkaaka oru kuRukiya paatai vazhiyaakac cenRaan
S + as (many) as + S I had eaten as (many) apples as Uma had eaten.	COMPARISON	NP…ettanai + NP + V- finite-oo + attanai + NP +… + V-finite umaa ettanai appiL caappiTTaaLoo attanai aappiL ndaanum caappiTTeen.
S + more than + S I had eaten apples more than Uma had eaten.	COMPARISON	*atu-ai* clause + *viTa kuuTutal*+ S umaa cappiTTatai viTa kuuTutal ndaan caappiTTeen
NP + BE + Adj (comparative degree) + than + NP I am taller than Usha.	COMPARISON	NP + NP-ai + viTa + Adj + pronominalizer + {-aay+iru} ndaan uSaavai viTa uyaramaanavan
S + although + S She thanked them although they refused her suggestion.	CONTRAST	adjectival clause + pootilum + S avarkaL avaL karuttai maRutta pootilum avaL avarkaLukku ndanRi kuuRindaaL.
S+ even though + S She behaved politely even though she was rich.	CONTRAST	adjectival clause + *pootilum* + S avaL paNakkaariyaay irundta pootilum avaL mariyaataiyaay ndaTandtu koNTaaL.
S + despite + S I ate two eggs despite (my) hating eggs.	CONTRAST	adjectival clause + *pootilum* + S Ndaan muTTaiyai veRutta pootilum iraNTu muTTaikaL caappiTTeen.
S + so (that) +S He had spent a lot of money on the project so (that) it was a success.	RESULT	V-infinitive + veeNTum + enpataRkaaka + S tiTTam veRRi peRaveeNTum enpataRkaaka paNam calavazhittaan.
Please get the letter from Raju if he had read it.	CONDITION	All-clause + S Raaju andta kaTitattaip paTittuviTTaal atai avaniTemirundtu vaangkit taravum.

The following are the different dependent clauses whose parallel structures are dealt here.

1. The nominal clauses
2. Adverbial clauses
3. Adjectival clauses
4. Comparative clauses
5. Coordiate clauses

4.2.1 Parallels in Nominal / Complement Clauses

Clauses in English	Clauses in Tamil
1.1. That – clause functioning as subject That she is beautiful is true.	1.1. S + *enpatu* + NP avaL azhakaanavaL enpatu uNmai
1.2. that – clause Functioing as direct object I told him that she was beautiful	1.2. S + *enRu* + S ex. avaL azhakaanavaL enRu avaniTam kuuRineen.
1.3. That – clause functioning as oppositive Your assumption, that things will improve, is understood.	S + *enRa* + S kaariyangKaL meenmaiyaTaiyum enRa unnuTaiya karuttu terikiRatu.
1.4. That – clause functioning as adjectival complement I am sure that things will improve.	S + *enRu* + S kaariyangkaL meenmaiyuRum enRu ndaan uRutiyaaka ndampukiReen.
2.1. Wh-clause functioning as subject What he is searching for is a house.	*atu* – clause + S avanm teeTikkoNTirukkiRatu oru viiTu
2.2. Wh – clause functioning as direct object He wants to eat whatever is ready.	Interrogative *oo*-clause + S etu tayaaraaka irakkiRatoo atai avan caappiTa virumpukiRaan
2.3. Wh – clause functioning as indirect object She gave whoever came a cup of	Interrogative oo-clause + S yaar vandtaarkaLoo avarkaLukku avaL teendiir koTuttaaL

tea.	
2.4. Wh – clause functioning as subject complement The truth is what I guessed correct.	Interrogative *oo*-clause + S etu cariyenRy ndaan ndinaitteenoo atu taan uNmai
3. If / whether – clause functioning as direct object I don't care if/whether he is available.	*aalum*-clause + S avan irundtaalum ndaan ataippaRRi kavalaip paTavillai
4. Nominal relative clause introduced by wh – element. What he is looking for is a wife.	*atu* -clause + S avan teeTikkoNTiruppatu oru manaiviyai
5.1. To – infinitive nominal clause functioning as subject For a boy to do that is strange.	*atu*-clause + NP oru paiyan itaic ceyvatu aticayamaanatu
5.2. To-infinitive nominal clause as subject complement Her intention is to become a doctor.	*atu* -clause + NP oru maruttuvaraavatu avaLatu viruppam.
6. Nominal ing – clause as subject complement. His hobby is collecting stamps.	*atu* -clause + S anjcal villaikaL ceekarippatu avanuTaiya pozhutu pookku
7. Bare infinitive All he did was press the button	atu – clause + S pottaanai amukkiyatu taan avan ceytatu
8. Verbless clauses Mosaic flooring in every room is expensive	*atu* -clause + S ovvoru aRaiyaiyum mucaik tarai pooTuvatu perunjcelavaakum.

4.2.2 Parallels in Adverbial Clauses

The following tables show the correlative features of adverbial clauses in English and Tamil.

Clause structures in English	Type of clause	Equivalent clause structures in Tamil
1.1. S + after + S I went to Madurai after	1. Clause of time	Adjectival clause + *piRaku, pinnaal, pinup, pinnar, pin*

she left Chennai.		avaL cennaiyai viTTu Poona pinnar ndaan maturai cenReen.
1.2. S + before + S She left the college before she finished her ecucation.	“	*ataRku*–clause + *munnaal, munpu, munner, mun* + S avaL tan paTippai muTippataRku munnaal kalluuruyai viTTup pooyviTTaaL.
1.3. S + Since + S I am working as teacher since we last met.	“	*atu* clause + *il* irundtu Poona taTavai ndaam candittatilirundtu ndaan aaciriyaraakap paNiyaaRRukiReen.
1.4. S + until + S I will be at Chennai until we meet again	“	Adjectival *um* – clause + *varai* + S ndaam miiNTum candtikkum varai ndaan cennaiyil iruppeen.
1.5. S + When + S I met you when you were in Chennai.	“	Adjectival clause + *pootu* ndii cennaiyil irundta pootu ndaan unnaic canditteen.
1.6. While… V-ing He watched them while they were eating	“	Adjectival clause (i.e.RC) + pootu avarkaL caappiTTukoNTirundta pootu avan avarkaLai kavanittaan.
2.1. S+ where + S I met her where she was working	2. Clause of place	… *engku* … V-finite – *oo angku* …. V-finite avaL engku veelai ceykiRaaloo angku avan avaLaic candtittaan. Or Adjectival clause + iTam avan avaL veelai ceykiRa iTattçk avaLaic candtittaan.
2.2. S +wherever + S He accompanied her wherever she went.	“	…*engkellaam..* V-finite oo + *angkellam*. V- finite. avaL engkellam cenRaaLoo angkellam avanum kuuTac cenRaan. Adjectival clause + iTanttiRkellaam avaL cenRa iTattiRkellam avanum kuuTac CenRaan.
3.1. If + S + S If he performs well, He will win the prize. If you wait, (then) You can meet him. 1. Real condition if she comes, I’ll talk to her.	3. Clause of condition “	*aal*-clause + S avan ndanRaaka ceyalpaTTaal, paricu peRuvaan. ndii kaattirundtaal avanaic candtikkalaam 1. Real condition avaL vandtaal ndaan avaLiTam peecuveen.
If it rains, the picnic will be cancelled.	“	mazhai vandtaal piknik rattu ceyyappaTum.

2. Unreal condition if she came, I'd talk to her.		2. Unreal condition avaL vandtirundtaal, ndaan peeciyiruppeen.
3.2. Unless + S S Unless it rains, the crops will die.	"	V + *aa* + viTTaal – clause + S Mazhai peyyaaviTTaal, payirkaL iRanduviTum.
3.3. Had + S + S Had I been there, I'd have congratulated you.	"	*aal* – clause + S ndaan angku irundtaal, unnai paaraaTTiyiruppeens
4.1. although + S + S Although he tried hard, he failed.	4. Clause of concession	Adjectial + *pootilum* – clause + S avan mikundta ciramappaTTa pootilum tooRRuviTTaan
4.2. if + S + S if he's poor, at least he's happy	"	Adjectival + *pootilum* – clause + S Avan ezhaiyaaka irundta pootilum makizhcci yaaka irukkiRaan
5.1. because + S I could not meet Uma, because she was ill.	5. Clause of reason or cause	Adjectival clause + kaaraNattaal + S umaa uTal ndalamillamal irundta kaaranataal ndaan avaLaic candtikka muTiyavillai.
5.2. S + since + S I did not work hard since you didn't give me salary.	"	Adjectival clause + *kaaraNattaal* ndii enakkuc campaLam taraata kaaraNattal ndaan kaTinamaaka uzhaikkavillai.
6.1. Since + S + S Since the rain has stopped, we shall go out.	6. Clause of circumstances	*aal*-clause + S Mazhai ndinRuviTTataal, ndaanm veLiyee poovoom.
6.2. As + S + S As it was dark, she hesitated to go out.	"	*aal*-clause + S iruTTaaka irundtaTaal, avaL veLiyee pooka tayangkinaaL
7.1. S + so that + S He worked hard, so that he could succeed.	7. Clause of purpose	U-clause enpataRkaaka + S veRRi peRaveeNTum enpataRkaaka avan kaTinamaaka uzhaittaan.
7.2. S + in order that +S We eat well, in order that we may be healthy.	"	*Um*-clause + enpataRkaaka raaju kuRRattai oppukkoLLaTTum enpataRakaaka kattirunteen
7.3. S + to-infinitive clause He walked fast, to catch the train	"	Infinitive clause + S avan rayilaip piTikka viraivaay ndaTandtaan
7.4. S + so as + infinitive – clause She studied hard. So as to get the scholarship	"	*veeNTum*-clause + *enpataRkaaka* + S avaL uukkattokai kiTaikka veeNtum enpataRkaaka ndanRaakap paTittaaL

7.5. S + in order + infinitive – clause He went through the narrow lane in order to avoid the police.	"	*VeeNTum* – clause + *enpataRkaaka* + S Ex. kavalariTamirundtu Tappa veeNTum enpataRkaaka paatai kuRukiya paatai vazhiyaakac cenRaan.
7.6. S + in order that +S We eat well, in order that we may be healthy	"	*veeNTum* – clause + *enpataRkaaka* aarookiyamaaka irukkaveeNTum enpataRkaaka ndaam ndanRaakac caappiTukiRoom
8. S + so that + S He practiced well, so that he could perform well. The dog barked so loudly that the thief fled	Clause of result	*um*-clause+*paTikku*+S avan taan ndanRaaka ceyalpaTum paTikku ndanRaaka payiRci ceytaan tiruTan ooTum paTikku ndaay kuraittatu
S + as if + S Raja acted as if he was sick	Similarity	*atu*-clause + *poola/pool* raajaa uTalndalamillaamal irundtatu poola ndaTittaan
S + as though + S He acted as though Usha has not seen him	Similarity	*atu*-clause + poola/pool uSaa avanaip paarkkaatatu poola avan ndaTittaan.
… as … as Ramesh came as frequently as Raja.	Similarity	NP ai + *viTa kuuTutal* raajaa umaavaivtha kuuTutai celavazhitt aaL
… more than + S Raja spent more than Uma had spent.	Difference	NP – ai + ViTa Raajaa rameeSai viTa viraivaaka ndaTandtaan
(al) though + S He gave them money although they refused to accept it.	"	Adjectival clause + *irundta pootilum* avaL paNakkaariyaaka irundta pootilum ndanRaaka ndaTandtukoNTaaL.
So (that) + S He spent lavishly so (that) he became pauper soon.	Reason	*ataRkaaka* – clause avan vaRiyavanaavataRkaaka taaraaLamaaka celavazhittaan

4.2.3 Parallels in Adjectival Clauses

Adjectival clause or relative clauses are clauses linked to a noun in their container clause, frequently with a WH form like the relative pronoun *which* and *whom.* As relative clauses qualify an NP, it performs the function of an adjective.

The girl who is clever

The following points have to be remembered while transferring a relative clause construction in English into Tamil.

1. In English the relative construction occurs in the finite from whereas in Tamil the verb in relative construction occurs in the non-finite form.
2. In English the verb follows the head noun whereas in Tamil, it precedes the head noun.
3. In both the languages, the verbal form in the relative constriction has time relation.
4. In English the relative pronoun has always a co referential noun whereas in Tamil, there is no co referential noun to the head noun.
5. In English, the relative construction occurs in conjunction to qualify the co referential noun whereas in Tamil, the relative participle occurs in succession to qualify the head noun.
6. In Tamil, an adjective may intervene between the non – finite verb and the head noun.

The following table gives the three types of relative clauses in English and their parallel clauses in Tamil.

Types of relative clauses in English	**Their equivalents in Tamil**
1. Restrictive relative clause as post modifiers: 1.1. [NP + [relative word +S]] NP The boy that is eating is her friend. 1.2. [NP + S] NP The table we bougt was strong.	[[…. V+ Tense /negative+relative participle]+ NP] NP caappiTTukkoNTiruppatu avaLuTaiya ndaNpan. Ndaam vaangkina meecai valimaivaayndtatu.
2. Non restrictive relative claise as post modifiers [[NP + [relative word +S]] NP The boy drawing the picture is my son.	[[… V+ Tense/negative + relative participle] + NP]] NP vaazhttu terivitta kaNNanai avaL candtittaaL
3. Non finite relative clause as post medifiers 3.1. (NP + [V-ing….]] NP The boy drawing the picture is my son	[[V+Tense/Negative + Relative Participle] + NP] NP paTattai varaindtu kiNTirukkiRatu ennuTaiya makan unnaaL taLLappaTTa manitar enmaamaa.

3.2. [[NP + [V3…]] NP The man rejected by you is my uncle. 3.3. [[NP [infinitive clause]] NP The next girl to sing is my daughter.	aTuttu paaTairukkum ciRumi enmakaL.

4.2.4 Parallels in comparative clauses

The following table correlates the comparative elements used in English and Tamil.

Meaning	Comparative elements in English	Comparative elements in Tamil	Comment
Similarity	as… as so … as She is as clever as her brother.	*pool*, *poola* avaL tan cakootaranaip poola aRivuLLavaL avaL tan cakootariyaip poola aRivuLLavalaay irukkiRaaL	Tamil makes use of pronominalized forms of adjectives instead of adjectives while denoting the present state/ quality. The formation can be captured by the following rule : [[Adj + [pronominalizer] NP + {aay]Adj+iru}]]
Dissimilarity	Than She is cleverer that her brother.	*vita, kaaTTilum* avaL tan cakootaranai viTa/ kaaTTilum aRivullavaL. avaL tan cakootariyai viTa/ kaaTTilum aRivuLLavaLLay	“

		irukkiRaaL	

4.2.4.1 Paralles in comparative clause of quality

The following table illustrates the transfer comparative clauses of quality.

Adjectival form in English	Degree	Parallel form in Tamil
NP + BE + as + positive form of adjective + as + NP Uma is as beautiful as Usha.	Similarity positive degree	NP + NP–*ai* + *poola* + Adjective + *iru* Umaa uSaavaip poola azhakaay irukki RaaL.
NP + BE + more + adjective + than + NP Uma is more beautiful than Usha.	Dissimilarity comparative degree	NP + NP – *ai* + *viTa* + Adjective + *iru* umaa uSaavai viTa azhakaay irukkiRaaL
NP + BE + Comparative form of adjective + than + all Raja is the tallest among all.	Dissimilarty comparative degree	NP + NP + *ai* + *viTa* + Adjective + iru ex.rajaa rameesai viTa uyaramaaka irukkiRaan.
NP + BE + the + Suforltive form of adjective + among all Raja is the tallest among all	Dissimilarity Super lative degree	NP + NP + elloorilum / ellaavaRRidam + Adjective + tiru raja avarkaL elloorilum uyaramaay irukkiRaan.

4.2.4.2 Paralles in comparative clause of quantity

The following table illustrates the transfer of comparative clauses of quantity.

Adjectival form in English	Degree	Parallel form in Tamil
… NP + HAS + as + many + NP + as … Ram has as many shirts as Sam has.	Positive degree clause of similarity	NP – <u>iTam</u> + ettanai + NP iru + Tense + PNG-oo + attanai + NP + NP – iTam iru + Tense + PNG RaamiTam ettanai caTTaikaL irukkinRanavoo attanai caTTaikaL caamiTamum irukkiRana.
NP… + HAVE + more +	Comparative degree	NP –iTam + ettanai + NP iru +

NP + than + NP + HAVE Ram has more shirts than san has.	Clause of difference	Tense + PNG-oo + atai viTa kuuTutal+NP+NP- iTam + iru + Tense + PNG RaamiTam ettanai caTTaikaL irukkinRanavoo atai viTa kuTutal caTTaikaL caamiTam irukkiRana

4.2.4.3 Paralles in comparative clause of adverbs

Adverbial comparative construction varies based on the three degrees of comparison. The three forms of adverbs, positive, comparative and superlative forms, can be referred from the DEWA.

English	Tamil
POSITIVE DEGREE With the positive form *as … as* in the affirmative and *as/so…* as in the negative are used Uma shouted as loudly as she could. It didn't cost her so much, as she feared.	umaa avaLaal muTindta aLavukku urakka captamilTTaaL avaL kavalaippaTTatu pool atu avaLukku KuuTutalaana vilai alla.
COMPARATIVE DEGREE With comparative form than is used. Uma walks faster than Usha. Kannan screamed louder than I expacted.	NP + NP – ai + ViTa + Adv + V Umaa uSaavai viTa veekamaaka ndaTandtaaL Ndaan etirpaarttatai viTa kaNNan kuukuraliTTaan
SUPERLATIVE DEGREE With superlative it is possible to use *of* + noun Usha worked hardest of the labourers.	NP + NP –il Adv + V uSaa veelaiyaaTkaLil kuuTutalaaka veelai ceytaaL.

4.2.5. Parallels in co-ordination

The following table depicts the points to be noted while correlating coordination in English to Tamil.

English	Tamil
In expressing coordination. English	Tamil as a SOV language, by contrast,

being an SVO language, place particles before the coordinated element, typically the last. Mathematics, physics, chemistry and zoology.	place such particles after the coordinated elements. kaNitamum, iyeRpiyalum, veetiyalum vilangkiyalum
Coordination is often accompanied by ellipses when two clauses are coordinated. Usha sat still and said nothing.	In this type of coordination, Tamil does not make use of the coordinator um. Intead it makes use sunordination by verbal participle form. uSaa acaiyaamal uTkaarndtukoNTu onRum peecavillai

4.3 Parallel structures of English and Tamil phrases

The following phrases are dealt with here.

1. Parallels in NP
2. Parallels in VP
3. Parallels in PP
4. Parallels in Adj P
5. Parallels in Adv P

4.3.1. **Parallels in noun phrases**

A typical noun phrase in English can be analysed as follows, which in turn can be transferred into Tamil by making use of the transfer rule.

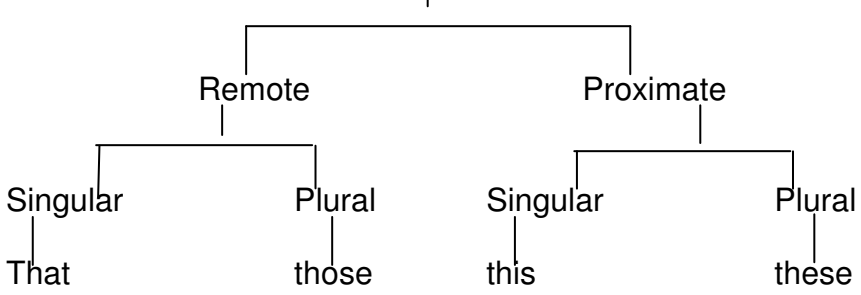
English: NP << Pre-det + Det + Ord + Quant + Adj P + Class + N
Tamil: NP << Class + Gen P + Qrd /Quant + Dem + Adj + N

The word order of the constituents in NP in Tamil is not fixed as in the case of English.

4.3.1.1 Prallels in demonstratives

Demonstratives in English and Tamil can be distinguished into two types in terms of proximity and remoteness. In English they can be further distinguished in terms of number. The following tree diagrams will illustrate the point:

Figure.1. **Demonstratives in English**



Demonstratives in Tamil

The following table gives the correlative features of English and Tamil demonstratives

English	**Tamil**
This	*indta* itu
That	*andta* atu
These	*indta* ivai
Those	*andta* avai

The following correlative features of determiners in English and Tamil have to be noted down while transferring NP in English into Tamil.

1. English have articles, which can be differentiated into definite and indefinite articles whereas Tamil has no article, which can be distinguished into definite or indefinite.
2. It is sometimes possible in Tamil to make use of demonstratives for the definite article.
3. English recognizes singular versus plural distinction under remote and proximate demonstratives whereas Tamil does not make such distinction.

4. English makes use of demonstratives as pronominals in certain places whereas Tamil does not make use of demonstratives as pronominals.

4.3.1.2 Parallels in quantifiers

The following points have to be noted down while transferring the quantifier system of English into Tamil.

1. The aggregates (*all, both,* etc.), fractionals (*half, one-third, two-third* etc.) and multiples (*Twice, three times, four times,* etc.) occur before determiners in English and so they are called pre determiners. But in Tamil, the aggregates (*ellaa* 'all', etc), fractionals (*paati* 'half' *muunRil onRu* 'one third', *muunRil iraNTu* 'two third', etc) and multiplicatives (*iraNTu maTangku* 'two times', *muunRu maTangku* 'three times', etc) do not precede the demonstratives (except in certain cases where it may be due to stylistic variation).
2. In English the prepositions either optionally or obligatorily links the head noun with certain set of quantifiers. Because of this reason fractionals and multiples in English occur before determiners, whereas in Tamil fractionals and multiples follow the determiners. The following examples will illustrate this point.

English	Tamil
all those girls	*andta ellaa ciRumikaLum*
all those things	*andta ella poruTkaLum*
two thirds of this portion	*Itil muunRil iraNTu pangku*
double the amount	*iraNTu maTangku tokai*
half of the crowd	*andta paati kuuTTam*
three fourths (of the) share	*andta mukkaal pakuti*

In Tamil the clitic *um* has to be added after the noun to bring in the aggregate meaning.

4.3.1.3 Parallels in genitive phrase

Genitive observes the arrangement of relative clause with regard to their head; a considerable number of genitive constructions in contemporary English do, follow their head.

The lap of her sister.

Her sister's lap.

If genitive, however, are proper nouns, particularly single names, they often precede noun.

John's house.

Yet even single names are often postposed.

House of John.

The current status of the genitive in English reflects a change form OV order to VO order. While the favoured order for genitives has been shifted, adjectives still predominantly precede the modified noun.

Large blue eyes.

Only when they are in turn modified do descriptive adjectives regularly follow their head.

They rested on a rock conveniently low.

Limiting adjectives–articles and demonstratives – also stand before nouns, as do numerals; they usually precede descriptive adjectives.

I could not hear him at that distance.

I haven't sent the two persons.

I jumped over the first of the six benches.

Parallel to the order of limiting adjectives is that of multiplying numeral combination with nouns representing higher entities millions, thousands, hundreds, tens in the order of higher to lower (preceded by the simple numerals).

Four thousand two hundred and seven.

As with proposed descriptive adjectives, genitives and relative clauses, proposed limiting adjectives and the cited numerals combinations reflect OV structure. This is the most conservative of the English modifying patterns. In maintaining it as a relic pattern, English provides evidence for the OV structure, which posited for its ancestor language.

Yet English nominal phrases for the most part observe the canonical order of SVO languages, maintaining form early stages OV order only with adjectives and numeral constructions other than the teens.

4.3.2 Parallel strucrues in Verb Phrase

In SVO languages, like English, expressions of verbal modification should be placed before verbs, in accordance with their VO structure. Like nominal modification, verbal modification avoids disruption of the VO constituent. The presence of separate verb like elements called auxiliaries constitutes one of the characteristics of SVO languages and of English. The expressions for declarative utterances simply observe the normal word order.

This speech caused a remarkable sensation among the party.

It is generally believed that Tamil lacks of VP constituency. So generally Tamil sentences are given a flat structure without VP being at a different hierarchical level. Tamil is an SOV language in which ‘S’ and ‘O’ can be shuffled. Tamil is not strictly a configurational language. The direct and indirect objects can be identified by case suffixes.

Many interesting points will be revealed for the purpose of transferring English language structure into Tamil, if we look at the correlating features of the two languages from the point of view of their typological characteristics as SVO and SOV languages respectively. Syntactically, English and Tamil are perhaps most saliently different in the basic word order of verb, subject, and object in simple declarative clauses. Tamil is an SOV language, meaning that the verb tends to come at the end of basic clauses. Tamil being SOV language has postpositions. Tamil is a typical (S) OV language in which the verb occurs at the final position of a sentence. Word

order in the sentence is relatively free, as long as the sentence ends with a main verb.

4.3.2.1 Parallels in complex verbal forms denoting tense, mood and aspect.

Both English and Tamil employ the complex process of combining both inflection and compounding in denoting tense, aspect, and mood. We can find correspondences between English and Tamil for the purpose of translating one from the other, though the correspondences are not always perfect; there are proper equivalents and improper and defective equivalents. The tense, aspect, and mood systems of English and Tamil operate differently and finding equivalents is a tough task. But for the purpose of MT, we compromise with certain peripheral differences between them and try to capture the core of the systems with the view in mind that what is conveyed in English can be transferred to Tamil without much distortions as our idea is to translate linguistic text in English into Tamil. The emotive and attitudinal senses conveyed by the auxiliary system will not play a vital part in expressing linguistic concepts. So we ignore the emotive and attitudinal sense and try to capture a core aspectual and model system. That is why we have ignored certain auxiliaries, which are used in Tamil to denote certain attitudinal and non-attitudinal senses. With this aim in mind, the aspectual and modals systems in both languages have been correlated for the purpose of preparing MTA. The following table correlates TAM system of English with that of Tamil.

TAM forms in English with examples	**Meaning**	**Equivalent TAM forms in Tamil with examples**
V1 + past tense He wrote	Past tense	V + past tense + PNG a*van ezhutinaan*
V + present tense He writes	Present Tense	V+present tense+PNG *avan ezhutukiraan*
has + have + V3 He has written. I have written	Present perfect aspect	V + past participle + *iru* + present+ PNG *avan ezhutiyirukki Raan.* ndaan ezhutiyirukkiReen
had + V3 He had written	Past perfect aspect	V + past participle + *iru* +past +PNG *avan ezhutiyirundtaan*
Be' verb +	Present progressive	V + past participle +

Present tense + V – in He is writing.	aspect	*koNTiru* + present - PNG *avan ezhutik koNTiukiRaan*
'Be' verb + past tense + V- ing He was writing	Past progressive aspect	V + past participle + *koNTiru* – past – PNG *avan ezhutik KoNTirundtaan*
Will/shall be verb future tense + V – ing He will be writing a letter.	Future progressive aspect	V + past participle + *KONTiru* + future + PNG *avan ka Titam ezhutik koNTiruppaan*

1.can + V1 He can speak English but he can't write it very well.	Ability = be able to = be capable of = know how to	V + infinitive + *mu Tiyum-iyalum* *avanaal aangkilam peeca muTiyum, aanaal ezhuta muTiyaatu.*
1. Can + V1 Can I smoke in here ('Am I allowed to smoke here?)	1. Permission = be allowed to = be permitted to ('Can' is less formal than 'may' in this sense)	1. V-*al + aam + aa* ndaan pukai piTikkalaamaa?
1.can + V1 Anybody can make mistakes. 2.can + be + V3 The road can be blocked.	1. Possibility = it is possible but / to theoretical possibility may = factual possibility	1. V-*al + aam* *yaarum tavaRu Ceyyalam* 2. V + infinitive + *paTal + aam* *caalai aTaikkppa Talaam.*
Could + V1 I never could play the chess.	Ability	V + infinitive + *muTiyum / iyalum* *enaal caturangkam aaTa muTiyavillai.*
Could + V1 Could I smoke in here?	II. Permission	V-*al + aam + aa* ndaan pukai PiTikkalaamaa?
1. Could be + C That could be my train. 2. Could be + V3 The road could be blocked.	II. Possibility (theoretical or factual, cf: might)	1.irukkal + *aam* Atu ennaTaiya to TarvaNTiyaay irukkalaam 2a. V + infinitive + *PaTTu* + irukkal + aam (aalai aTaikkap PaTTu irukkalaam 2b. V + infinitive + *PaTTu* + irukka + kuuTum (aalai aTaikka PaTTirukkak kuuTum
may + V1 He might leave tomorrow	Future time with modal auxiliaries. In many contexts, modal	V-*al + aam* avan ndaalai Pookalam

	auxiliaries have inherent further reference, both in their present tense and past form.	
1.may + V1 You may borrow Car if you like. 2. may not + V1 ex. You may not borrow my car (=You are not allowed to borrow my car.)	III. Permission =be allowed to = be permitted to in this sense may is more formal than can. Instead of may not or mayn't mustn't is often used in the negative to express Prohibition	1.V-al + aam ndii ennuTaiya kaarai KaTanaakap peRalaam 2.V+infinitive + maaTTu +PNG .ndii ennuTaiya kaarai KaTanaakap PeRamaaTTaay
1.may +V1 He may never Succeed ('It is Possible that he will never succeed') 2. may + be + V_3	III. Possibility = be it is possible that /to May – factual possibility (cf: can = theoretical possibility)	1a. V-*al* +*aam* (for positive meaning) 1.b. V+ infinitive + *maaTTu* + PNG avan veRRi peRamaaTTaan 2. V+ infinitive +paTal +aam Caalai aTaikkappaTalaam.
might + V1 He might leave tomorrow.	Future time with modal auxiliaries. In many contexts, modal auxiliaries have inherent future reference, both in their present tense and past for.	V-*al*+*amm* avan ndaalai pookalaam
Might…? Might I smoke here?	IV. Permission	V-*al* + *aam* + *aa* Ndaan pukai PiTikkalaamaa?
Might + V1 He might succeed.	Possibility	V-*al* + *aam* Avan veRRi peRalaam.
Shall + V1 He shall get the money. You shall do exactly as you wish.	II. Willingness on the part of the speaker in 2nd person and 3rd person ('weak volition') Restricted use.	V-al + aam Avan paNam peRalaam. Ndii virumpuvatu pool Ceyyalaam.
Shall + V1 We shall let you know our decision. We shall, overcome. 1b. Shan't + V1 It shan't be long for me to meet the minister.	1. Intention on the part of the speaker only in 1st person ('intermediate volition')	1a.V+future tense +PNG ndaagkaL ungkaLiTam engkaL tiirmaanattait terivippoom. 1b. aakaatu enakku mandtiriyai paarkka ndiiNTa ndeeram

		aakaatu
Shall + V1 You shall do as I say. He shall be punished. The vendor shall Maintain the equipment in good repair.	1.a. Insistence ('Strong volition'). Restricted use. b. Legal and quasi-legal.	V-al+aam Ndaan collukiRa paTi Ceyyalaam. Avan taNTikkappa Talaam. Teru viyaapaaari tannuTaiya Karuviyai ceppam ceytu Vaittirukkalaam.
Should + V1 You should do as he says. They should be home by now.	1.Obligation and logical necessity (=ought to)	V+infinitive + veeNTum ndii avan colvatu pool Ceyya veeNTum avarkaL ippootu viiTTil irukka veeNTum.
Will/shall + v1 He will write. I shall write.	Future tense	V + future + PNG avan ezhutuvaan Ndaan ezhutuveen
Will + V1 I'll write as soon as I can. Will you have another cup of tea?	I. Willingness ('weak volition') unstressed, especially 2nd person. "Down toners' like please may be used to soften the tone in requests.	V+future tense + PNG Epootu muTikiRatoo apootu Ndaan ezhutuveen. Innoru kooppai teenier kuTikkiRaayaa?
Will + V1 I'll write as soon as I can. We won't stay longer than two hours.	II. Intention (intermediate volition). Usually contracted 'II', mainly 1st person.	V+future tense +PNG muTiyumpootu uTanee ezhutuveen. ndaangkaL iraNTu maNi ndeerattiRku meel tangka maaTToom.
Will + V1 He 'will do it, whatever you say ('He insists on doing it…') (cf He 'shall do it, whatever you say = 'I insist on his doing it')	II. Insistence ('strong volition'= insist on). Stress = ed, hence on 'll contra-diction. An un-common meaning.	V+future tense + PN G ndiingkaL colvatai avan ceyvaan.
would….? Would you excuse me?	III. Willingness ('Weak volition')	V+future tense + PNG + aa ndii ennai Mannippaayaa?
Would + V1 It's you own fault: you 'would take the baby with you.	III. Insistence ('Strong volition')	V+ past participle form + Iru + infinitive + veeNTum atu un tavaRu. Ndii Kuzhandtaiyai unnuTan eTuttuc cenRirukka veeNTum
1.must You must be back by 10	II. Obligation or compulsion in the	1.V+infinitive +veeNTum ndii pattu maNikku

o'clock. 2.had to Yesterday you had to be back by 10 o'clock. Yesterday you said you had to / must be back by 10 o'clock.	present tense = (=be obliged to, have to); except in reported speech. Only had to (not must) is used in the past.n the negative sentence needn't, don't have to, not be obliged to are used (but not must not, mustn't which = 'not be allowed to')	tirumpa veeNTum. 2.V + past participle +iru+infinitive+veeNTum ndeeRRu ndii condaay pattu maNikku ndaan tirumpiyirukka veeNTum enRu.
Will, must, Should + V1 The game will/must/should be finished by now.	[Prediction of the similar meanings of other expressions for logical necessity and habitual present. The contracted form 'll is common]. Specific prediction.	V+infinitive + veeNTum viLaiyaaTTu inndeeram muTindtirukka veeNTum.
Will + V1, V1 Oil will float/floats on water.	Timeless Prediction	V + future + PNG eNNai taNNiiril mitakkum.
Will, 'll He'll (always) talk for hours if you give him the chance.	Habitual prediction	V+future +PNG Avanukkuc candtarppam koTuttaal avan (eppozhutum) maNiokkaNakkaakap peecuvaan.
1.must, has to There must / has to be a mistake. (must is not used in sentences with negative or interrogative meanings, can being used instead .	Logical necessity.	1. irundirukka + veeNTum 2. tavaRu irundtirukka 3. veeNTum
2.cannot There cannot be a mistake. 3. mustn't (Must can occur superficially interrogative but answer- assuming sentence) Mustn't there be another reason for his behaviour.		2. irudndirukka + muTiyaatu tavaRu irundtirukka muTiyaatu. 3. irundirukka + veeNTaam avanuTaiya parimaaRRattiRku veeRu kaaraNam irundirukka veeNTaam.
ought to + V1 You ought to start at once. They ought to be	Obligation and logical necessity.	V+infinitive +veeNTum Ndii uTanee puRappaTa veeNTum.

here by now.		avarkaL ipootu ingkee irukka veeNTum.
Used to + V1 He used to fish for hours. He used to be an excellent cricketer.	A state of habit that existed in the past but has ceased. (cf: would, and formerly or once + past)	V-atu+vazhakkam +aay + iru + past + PNG maNikkaNakkil miin piTippatu avan vazhakkamaay irundtatu. Avan oru arumaiyaana Kirikket aaTTakaaranaaka Irundtaan.

The following points have to be noted while transferring TAM system of English into Tamil.

1. Both English and Tamil make use of inflection as well as compounding (i.e. combining main verbs with the auxiliary verbs) to express TAM.
2. The important point to be noted from the point of view of word order is that auxiliary verbs in English precede the main verb in English, whereas in Tamil they follow the main verb.
3. In English auxiliary verbs are inflected for tense, person and number; whereas in Tamil they are inflected for tense, person, number and gender.
4. Both English and Tamil undergo auxiliary reduction.
5. Identical auxiliary verbs in complex constructions get deleted both in English and Tamil.
6. Auxiliary verb occurs in a sequence to denote tense, mood, aspect, voice etc. In both English and Tamil. The modal auxiliary verb in English never occurs after a primary auxiliary verb in English, whereas; in Tamil primary auxiliary verb never occur after a modal auxiliary verb (with the exception of few modal auxiliary verbs).
7. Both Tamil and English express perceptive and progressive sense by auxiliary verbs. But Tamil makes use of this device to express the completive and reflexive sense too.
8. In English, interrogative sentences are derived from their respective affirmative sentences by shifting the relevant auxiliary verbs to the initial position.

The following table shows the correlative features of auxiliary system in English and Tamil for the sake of transfer.

Auxiliary system in English	**Auxiliary system in Tamil**
The auxiliary verbs are used to add auxiliary meaning to the main verb.	In Tamil too, a lot of auxiliary verbs are used to add auxiliary meaning to the main verb.
The auxiliary verb precedes the main verb.	The auxiliary verb follows the main verb.
Primary auxiliary verbs are inflected for tense, person and number.	Primary auxiliary verbs are inflected for tense, person, and gender.
The modal auxiliary verbs are inflected for negation.	Tamil too, the modal auxiliary verbs are inflected for negation.
-	The main verbs in the auxiliary verb constructions occur in the form of verbal participles infinitives or verbal nouns.
-	Any one of the identical auxiliary verbs is deleted in the formation of compound sentences.
Auxiliary verb reduction is possible.	Auxiliary verb reduction is possible in Tamil too.
Lexical insertion between the main verb and auxiliary verb (with the exception of a few modal auxiliary verbs) is possible.	Lexical insertion mentioned in the case of English is not possible in Tamil.
Auxiliary verb occurs in a sequence to denote a different grammatical meaning.	This is the case with Tamil too.
A modal auxiliary verb never occurs after a primary verb	A primary auxiliary verb never occurs after a modal auxiliary verb.
A statement can be converted into a yes – or – no question by shifting the auxiliary verb to the sentence initial position.	-
Auxiliary verbs occur in tag questions, but the main verbs do not occur in tag questions.	This is the case with Tamil too.
The *n't* that is the contracted form of *not* is cliticized onto the preceding auxiliary verbs but never onto a preceding main verb.	-
The supportive *do* appears before a main verb in certain environments, but never before an auxiliary verb.	-

The *all* (quantifier), which follows the auxiliary verb, is semantically associated with the subject of the sentence.	-

4.3.2.2 Paralles in verb patterns

The following table illustrates the transfer of verb pattern forms in English into Tamil.

Verb patterns in English	Parallel patterns in Tamil
VERB PATTERN 1 NP + BE + NP This is a book	 NP + NP atu our puttakam
NP + BE + PN It's me	NP + NP atu ndaan / atu ennu Taiyatu
NP + BE + Possessive P That's mine	NP + Possessive Pronoun – uTaiyatu atu ennuTaiyatu
Interrogative + BE + NP Who is that?	NP + interrogative pronoun atu yaar?
NP + BE + Adj. She is clever	NP + Adj – pronominalizer aval azhakaanaval
NP + BE + Adj.P The statue will be life size	NP + Adj – pronominalizer Andta cilia mikapp periyathu
NP + BE + PP She is in good health	NP + NP-ooTu + iru avaL ndalla aarookkiyattooTu irukkiRaalL
NP + BE + Adv She is here	NP + Adv + iru avaL ingke irukkiRaaL
There + BE + NP There was a large crowd	Angkee + NP + iru angkee perum kuuTTamaaka irundtatu
There + BE + NP + PP There are three windows in this room	PP + NP + iru Indta aRaiyil muunRu jannalkaL irukkiRana
It + mBE + Adj. / NP + to – infinitive It is so nice to sit here with you.	atu-clause + Adj. + iru unnuTan uTkaarndtu iruppatu rompa ndanRaaka irukkiRatu
How + Adj./NP + (it+BE) to – infinitive phrase How nice it is to sit here with you	atu-clause + evvaLavu + Adj + iru unnuTan uTkaarndtu iruppatu evvaLavu ndanRaaka irukkiRatu
What + Adj. / NP + (it + BE) to infinitive clause What a pity it is to waste time.	atu-clause + ervvaLavu + Adj + iru ndeerattai viiNaakkuvatru evvaLavu moocamaaka irukkiRatu

It BE + Adj. / NP + gerundial clause It is so nice sitting here with you	atu-clause + Adj. + iru UnnuTan UTkaarndtu iruppatu ndanRaaka irukkiRatu
NP + BE + that-clause The trouble is (that) all the shops are shut.	enpatu- clause + Adj + iru ellaa kaTaikaLum muuTappaTTirukkinRana enpatu cikkalaaka irukkiRatu
It + BE + NP / Adj. + that – clause It was a pity (that) you couldn't come	enpatu-clause + Adj + iru unnaal vara iyalavillai enpatu varattattiRkuriyataaka irundatu
NP + BE + to-infinitive clause This house is to let It + BE + Adj. / NP + for + N/ Pronoun + to-infinitive (phrase) It was hard for him to live in this small cell.	NP + infinite-clause + iru Indta viiTu vaaTakaikku vita irukkiRatu atu-clause + Adj + iru indta cinna ciRaiyil vaazhvatu kaTinamaaka irundtatu
VERB PATTERN 2 NP + Vi The sun is shining	Cuuriyan pirakaacikkiRatu NP + Vi
There + vi + NP There followed a long period of peace and prosperity.	NP + vi Amaitiyaana vaLamaana Kaalam toTarndtau
It + Vi + whether –clause It does not matter whether we start now or latter.	Enpatu-clause + Vi ippozhutaa allatu pinnaraa ndaam toTangakvirukkiRoom enpatu viSayam alla .
It + Vi + to-infinitive clause It only remains to wish you both be happiness.	Infinitive clause + Vi ndiingkaL iruvarum makizhcciyaaka irukka vaazhtta irukkiRatu
It + Vi + that-clause It seemed that the day would never end.	enRu – clause + Vi ndaaL muTivuRaatu enRu toonRukiRatu
NP + Vi + for + PP (adv. Adjunct) We walked for five miles	NP + NP (Adv adjunct) + Vi Ndaam aindtu maikaL ndaTandoom
NP + Vi + Adj. Adjunct My hat blew off	NP + Vi ennuTaiya toppi PaRandtu viTTatu
NP + Vi + Adj. The leaves are turning brown.	NP + Adj + Vi ilaikaL pazhuppu ndiRattil maaRukinRana
NP + Vi + Adjectival past participle You look tired	NP + Adj + Vi Ndii kalaippaakat toonRukiRaay
NP + Vi + NP	NP + Adj + Vi

He died a millionaire	Avan oru laTcaatipatiyaaka iRandtoan
NP + Vi + Reflexive pronoun You are not looking yourself today	NP + Pronoun – aaka + Vi Ndii inRu ndiiyaakat toonRavillai
NP + Vi + V-ing + infinitive-clause The children came running to meet us.	NP + infinitive-clause + V-past participle + Vi kuzhandaikaL ndammaic can dtikka ooTivandtana
VERB PATTERN 3 NP + Vi + preposition + NP (NP= noun, pronoun) You can rely on me	NP + NP + postposition + vi NP = noun, pronoun ndii en militu ndammpikkai vaikkalaam
NP + Vi + preposition + NP + to-Infinitive Clause They advertised for a young girl to look after the children.	NP + infinitive-clause + NP + postposition + Vi avarkaL kuzhandtaikaLai kavanikka veeNTi oru iLamaiyaana peNNiRku viLamparam ceytaarkaL
NP + Vi + (preposition + (it) + that-clause We will see (to it) that she gets home early. They decided (on) who should act as Sita.	Infinitive – clause + NP + Vi avaL kaalam taazhttaamal viiTTiRku vara ndaam eeRpaaTu ceyvoom. Infinitive-clause + NP + Vi avarkaL muTivueTuttu viTTaarkaL yaar seettaavaaka ndaTippatu enRu.
VERB PATTERN 4 NP + Vi + to-infinitive clause (of purpose, outcome, or result) He ran to chase the thief.	NP + infinitive clause + Vi avan tiruTanait turatta ooTinaan.
NP + Vi + to-infinitive clause (may be equivalent to a coordinated or subordinated clause) He turned to see the sun setting.	NP + infinitive-clause + Vi avan cuuriyan maRaivataip paarkkat tirumpinaan
NP + Vi + to-infinitive clause (Infinitive adjunct is used after some verbs) She agreed to sing a song	NP + infinitive-clause + Vi avaL oru paaTTupaaTa cammatittaaL
NP + seem/appear + (to be) + Adj./NP This seems (to be) light thing NP + seem / appear + (to be) + Adj. /NP She seemed (to be) unable to enjoy it.	NP + Adj/NP-aaka + toonRu/teri atu ileecaana poruLaakat toonRukiRatu enRu-clause + toonRu/teri avaL atai iracikka iyalaatu enRu toonRukiRatu.
NP + be + to-infinitive clause You are to break the news	NP + infinitive clause + veeNTum ndii anda ceytiyai veLippaTutta

	veeNTum
VERB PATTERN 5 NP + Anomalous finites + bare infinitives You may leave now.	NP + infinitive clause + aam Ndii ippootu pookalaam
VERB PATTERN 6 NP + Vt + NP (passivisation is possible) Nobody answered my question.	NP + NP-kku + Vt Yaarum ennuTaiya keeLvikku viTai aLikkavillai
NP + Vt + NP (passivisation is not possible) She laughed a merry laugh.	NP + NP + Vt avaL inimaiyaana cirippu cirittaaL
NP +Vt+Gerundial clause (not replaceable by to infinitive She enjoys playing tennis	NP + infinitive- Clause + Vt avaL Tenni ViLaiyaaTa VirumpukiRaaL
NP + Vt + Gerundial clause (Replaceable by to-infinitive clause) The childrem love playing on the sand.	NP + atu-clause-ai+Vt kuzhandtaikaL maNalil viLaiyaaTuvatai virumpukinRana
NP+ need/want/bare + Gerund (With passive meaning) The garden needs watering	Atu-clause + NP tooTTAttil taNNiir viTuvatu avaciyam.
VERB PATTERN 7 NP+Vt + (not) + to-infinitive clause I prefer not to start early	Atu-clause-ai + Vt Munnaree toTangkuvatai ndaan virumpavillai
NP + have/ought + (not) + to-infinitive clause You ought not to complain	NP + infinitive-clause + kuuTaatu ndii kuRaikuuRak kuuTaatu
VERB PATTERN 8 NP + Vt + interrogative Pronoun + to-infinitive clause I don't know who to go for advice	NP + interrogative pronoun-iTam + atu-clause + enRu + Vt Ndaan yaariTam aRivuraikkaaka Poovatu enRu teriyavillai
She cound not decide what to do next	NP + interrogative pronoun-iTam + atu-clause + enRu + Vt avaLukku aTuttu enna ceyvatu enRu muTivu ceya iyalavillai
NP + Vt + interrogative Adv. + to-infinitive Clause Have you settled where to go for your holidays?	Interrogative Adv+V-atu+enRu+NP+ Vt Un ViTumuRai ndaaTkaLil en… PooVatu enRu ndii tiiirmaanittu ViTTaayaa?
NP + Vt + whether + to- Infinitive Clause She didn't know whether to cry or to laugh	atu-clause-aa enRu + NP-ukku + Vt taan cirippataa azhuvataa enRu avaLukkut teriyavillai

VERB PATTERN 9 NP + Vt + that-clause He doesn't believe that my intentions are serious.	atu-clause + enRu + NP + Vt ennuTaiya viruppangkaL mukkiyattuvam vaayndtatu enRu avan ndampavillai
VERB PATTERN 10 NP + Vt + wh-clause I don't know who she is. I don't know who is she.	[NP + interrogative word + (V)]S + enRu + NP-ukku + Vt yaar avaL enRu enakkut teriyaatu avaL yaar enRu enakkut teriyaatu
NP + Vt + whether-clause She asked whether I put sugar in my tea.	S-aa + enRu + NP +Vt Ndaan teendiiril carkarai pooTeenaa enRu avaL keeTTa
VERB PATTERN 11 NP + Vt + NP + that-clause He warned us that the roads were covered with snow.	S + enRi + NP + NP-ai + Vt caalaikaL panikkaTTikaLaal muuTappaTTuLLana enRu avan engkaLai eccarittaan.
VERB PATTERN 12 NP + Vt + NP [IO] +NP [O] The indirect object can be coverted into *to* + NP He handed her the letter (= He handed the letter to her)	NP + NP-iTam + NP + Vt avan avaLiTam kaTitaittai oppaTaittaan
NP + Vt + NP [IO] + NP [O] The indirect object can be converted into for + NP Are you going to buy me Some? (=Are you going to buy some for me?)	NP + NP-aaka + NP + Vt ndii enakku-aakak konjcam vaangkap pookiRaayaa?
NP + Vt + NP + NP He struck the door a heavy blow. .	NP + NP-aaka + NP + Vt Avan Katavukk oru Veeka maanatattu KoTuttaan
VERB PATTERN 13 NP + Vt +NP [O] + to + NP She read the letter to all her friends	NP + NP-iTam + NP-al + Vt avaL tannuTaiya …… cineekitikaLiTamum andta KaTitattaip PaTittuk'kaaTTinaL
NP + V + NP [O]+ for + NP I have bought some Cholocate for her	NP + NP-aaka + NP + V + Ndaan avaLukkaaka Konjam CaakleeTT VaankigirukkiReen
VERB PATTERN - 14 NP + Vt + NP [O] + for + NP We congratulated him on his success. Thank you for your kind help.	NP + NP-aaka + NP-ai + Vt. ndaangkaL avan veRRikkaaka avanaip paaraaTTincom. ungkaLuTaiya anpaana utavikku ndanRi
NP + Vt + PP + NP [O] I explained to him the impossibility of	NP + NP-ai + NP-iTam + Vt avanuTaiya veeNTukooLukku

granting his request.	utavipuriya iyalaamaiyai ndaan avaniTam viLakkineen
NP + Vt + PP + to-infinitive + that clause I must leave it to your own judgment to decide whether you should offer your resignation.	enRu clause +atu-clause-ai + NP + NP-ukku + Vt. ndii unnuTaiya iraajinaamaavai koTukkaveeNTumaa enRu muTivu ceyvatai ndaan unnuTaiya tiirmaanattiRkee viTTuviTaveeNTum.
VERB PATTERN 15 NP + Vt + NP [O] + Adv. P/PP Please put the milk in the refrigerator.	NP + NP-ai + NP-iL/Adv + Vt tayavu ceytu paalai kuLircaatanap peTTiyil vai.
NP + Vt + NP [O] + Adverbial Particle Put your shoes on	NP + NP-ai + VtT unnuTaiya kaalaNikaLai aNi
VERB PATTERN 16 NP + Vt + NP [O] + to-infinitive clause He brought his brother to see me.	NP+ NP-ai + infinitive-clause + Vt. avan tannuTaiya cakootaranai ennaip paarkka koNTu vandtaan
NP + Vt +NP [O] + as/like/for + NP They have hired a fool as our football coach.	NP + NP-ai + NP-aaka + Vt avarkaL oru muTTaaLai engkaLuTaiya kaal pandtaaTTA payiRciyaaLaraaka vaaTakaikku eTuttirundtaarkaL.
VERB PATTERN 17 NP + Vt + NP [O] + (not) + to-infi itive clause. (Can be passivized) I warn you not to believe a word he says.	[…. V-infinitive + veeNTaam]S + enRu +NP +NP-ai +Vt avan kuuRuvatil oru collaikkuuTa ndampa veeNTaam enRu ndaan unnai eccarikkiReen.
NP + Vt + NP + (not) + to-infinitive clause. (Cannot be passivized) He doesn't want anyone to know that she is going away.	atu-clause-ai + NP +Vt avaL veLiyeeRikkoNTirukkiRaaL enRu yaarum aRivatai avan virumpavillai.
VERB PATTERN 18 NP + Vt + NP + infinitive clause (Verbs indicate physical perception) Did any one hear John leave the house? Did you see anyone go out?	Atu-clause-ai + NP + Vt jaaN viiTTai viTTu veLiyeeRiyatai yaaraavatu paarttiirkaLaa? Ndii yaaraavatu veliyee poonataip paarttaayaa?
NP + Vt + NP + infinitive clause (Verbs do not indicate physical perception) What makes you think so?	NP + NP-ai + infinitive-clause + Vt etu unnai avvaaRu eNNat tuuNTiyatu?
NP + have + NP + infinitive clause	Atu-clause + ukku + …+Vt

We have computers to do our work.	ndaangkal engkaL veelaiyai ceyvataRku kaNini vaittuirukkiroom
VERB PATTERN 19 NP + Vt + NP + ing-clause (Verbs indicate physical perception) He felt his heart throbbing.	NP + atu-clause-ai + Vt Avan tan itayam aTippatai uNarndtaan.
NP + Vt + NP + ing-clause (Verbs do not indicate the Physical Perception) I can't have you doing that…	NP + NP-ai verbal participle Clause +….. + Vt Ndaan atai Unnai Vaittu Ceyya iyalaaku.
NP + Vt + NP + ing-clause (NP = noun, pronoun, possessive) I can't understand him/his Leaving so suddenly.	NP + atu-clause-ai + NP/NP-aal + Vt avan tiTiir enRu veLiyeeRuvatai purindtukoLLa iyalavillai
Verb pattern 20 NP + Vt + NP + interrogative on noun or adverb + to-infinitive clause I showed them how to do it.	[interrogative pronoun/Adv + V- atu] + enRu + NP + NP-ukku + Vt. evvaaRu ceyvatu enRu ndaan avarkaLukku kaaTTineen
NP + Vt + NP + whether + to-infinitive clause Ask her whether to trust him or not.	atu-clause-aa + enRu + NP + NP-iTam + Vt avanai ndampuvataa veeNTaamaa enRu avaLiTam keel
VERB PATTERN 21 NP + Vt + NP + wh-clause Tell me what your name is? He told me why he had come.	[…. Wh-word …] enRu + NP + NP-iTam + Vt un peyar enna enRu enniTam kuuRu avan een vandtaan enRu enniTam connaan.
VERB PATTERN 22 NP + Vt + NP [O] + Adj. (NP = noun, pronoun, gerund) We painted the ceiling green. The blister on my heal made walking painful.	NP + NP-kku + Noun of quality equivalent to adjective + Vt ndaangkaL uTkuuraikku paccai varNan aTittoom NP + NP-ai + Adj + Vt en paatattil irunda puNdaTappatai veetanaikkuriyataay ceykinRatu.

VERB PATTERN 23 NP + Vt + NP [O] + NP [Object complement] The team has voted me their new captain.	NP + NP-ai + NP-aaka + Vt andta kuzhu ennai avarkaLin putiya talaivanaaka teerndteTuttirukkinRatu
NP + Vt + NP + NP (Subject complement)	NP + NP-ai + NP-aaka + Vt Jill jaakkiRku oru ndalla

Jill has made jack an excellent wife.	Manaiviyaaka uruvaakiyirukkiRaaL
VERB PATTERN 24 NP + Vt + NP [O] + Past Participle Phrase NP [O] = noun, pronoun You must get this door painted You must make yourself respected	NP + [NP-ai+Infinitive] infinitive-clause + Vt Ndii indta katavai varnam Puucac Ceyya VeeNTum NP + [NP-uuku+V-infinitive] infinitive – clause + Vt. Ndii unaker mariyaatai tarac ceyya VeeNTum
NP + Vt + NP [O] + Past participle phrase NP [O] = noun, pronoun She's had her handbag stolen. The pilot had his plane hijacked.	NP + NP-ai + V-infinitive + Vt avaL tannuTaiya kaipaiyai tiruTa viTTaaL vimaanam ooTTiyee tan vimaanattai kaTattac ceytaar.
NP + have/get + NP [O] + Past participle NP [O] = noun, pronoun Please get the machine Repaired.	NP + NP-ai + V-infinitive + Vt tayavuceytu poRiyai pazhutu Paarkkac cey.
VERB PATTERN 25 NP+Vt + NP [O] + (to be) + Adj./NP Most people considered him (to be) innocent. I have always found Jonathan a good friend.	NP + NP-ai + NP-aaka + Vt palar avanai ndiraparaatiyaakak karutinaar. ndaan joonaatanai oru ndalla ndaNpanaakak KaNTeen.

4.3.3 Parallels in adjectival phrases

The following points of typological correlation have to be noted while attempting to transfer adjective phrase in English into Tamil.

1. Adjectives precede the nouns, which they qualify in Tamil, which reflect the characteristic of SOV language. Adjectives precede the nouns they qualify even though English is an SOV. However, a few numeral adjectives and all predicative adjectives follow the nouns, which they qualify. This tendency reflects the transition of sentence structure form SOV to SVO.

2. In English a few simple adjectives are inflected for degrees of comparison. Whereas in Tamil the adjectives are not inflected for degrees of comparison.

3. Adjectives occur in succession as qualifiers of head in both languages.

The following table illustrates the transfer of adjectival patterns in English into Tamil.

Adjectival Patterns of English	Parallel Patterns in Tamil
[… + Adj. + N] NP + V A good boy came	[… Adj + N] NP + V oru ndalla paiyan vandtaan
NP + BE + [… +Adj. + N] NP He is a good boy	NP + [… Adj. + N] NP avan oru ndalla paiyam
NP + BE + Adj. She is beautiful	NP + Adj + BE avaL azhakaay irukkiRaaL
NP + BE + Intensifier [Adv] + Adj. She is very beautiful.	NP + intensifier + Adj + Adj + BE avaL mika azhakaay irukkiRaaL
It + BE + Adj. + to-infinitive clause It's easy to please Jim	NP + atu-clause + Adj + BE Jimmai tirupptipaTuttuvatu eLitaaka irukkiRatu.
NP + BE + Adj. + to-infinitive clause Jim is eager to please every one	NP + infinitive clause + Adj + BE Jim ovvoruvaraiyum tiruptipaTutta Viruppamaaka irukkiRaan
It + BE + Adj. + to-infinitive clause It is wrong of Jim to leave	NP + atu-clause + Adj + BE jim veLiyeeRuvatu tavaRaaka irukkum
It + BE + Adj. + that-clause It is certain that Jim will win.	enpatu-clause + Adj. BE jim jeyippaan enpatu ndiccayamaaka irukkiRatu
It + BE + Adj. + to-infinitive clause John was first to arrive	Adv + V-atu + NP Mutalil vandtu cerndtatu jaaN
NP + BE + Adj. + (Preposition + NP] PP John is anxious for news NP + Adj. + (+preposition) + Clause John is glad that you succeeded. John is anxious about how they got on.	NP + [NP-ai + postposition] PP + Adj + iru jaaN ceytiyaip paRRi kavalaiyaaka irukkiRaan. NP + enRu-clause + NP + Adj + iru ndii veRRi peRRaay enRu jaaNukku mazhcciyaaka irukkiRatu. avarkaL eppaTi camaaLikkiRaarkaL enRu jaaNukku kavalaiyaaka irukkiRatu.

4.3.4 Parallels in Adverbial Phrase

The following points of typological comparison have to be noted while attempting to transfer adverbial phrases in English into Tamil.

1. An adverb occurs in attributive construction with a verb, an adjective, an adverb or a main clause in both English and Tamil.

2. Adverbs in English are inflected for degree of comparison whereas adverbs in Tamil are not inflected for degrees of comparison.

3. Adverbs in English follow the forms, which they modify as in other SVO languages whereas in Tamil they generally precede the forms, which they modify.

4. More than one adverb can occur in a sequence in both English and Tamil.

5. In Tamil adverbial stems are used in repetition to give more emphasis to the meaning expressed.

Umaa miiNTum miiNTum vandtaaL

'Uma came again and again'

The following table illustrates the transfer of adverbial patterns in English into Tamil.

Sno	**Adverbial Patterns of English**	**Parallel Patterns in Tamil**
	Adverbs of Manner	
1	NP + Vi + Adv.1	NP + Adv + Vi
	Usha ran fast.	avaL veekamaaka ooTinaaL
2	NP + Vt + NP [O] + Adv 1	NP + NP-ai + Adv. + Vt
	I ate banana hurriedly.	avaL vaazhaippazhattai viraivaaka caappiTTaaL
3	NP + Adv1 + Vt +NP [O]	NP + NP-ai + Adv + Vt
	Usha warmly welcomed the minister from Chennai.	uSaa cennaiyilurundtu vandta mandtiriyai anpooTu varaveeRRaaL
4a	NP + Adv1 + V + to-infinitive clause.	NP + Adv + infinitive-clause + V
	They secretly decided to go to Chennai.	avarkaL irakaciyamaaka cennai cella muTivu ceytanar.
4b	NP + V + to-infinitive clause + Adv1	NP + infinitive-clause + Adv + V

	They decided to go to Chennai secretly	avarkaL cennai cella irakaciyamaaka muTivu ceytanar.
5a	NP + V + NP [O] + Adv.11 (ex. Foolishly, generously, etc.,)	NP + Adv. + NP-ukku + vt
	Usha answered the question foolishly	uSaa muTTaaLtanamaaka keeLvikku viTaiyaLittaaL
5b	NP + Adv1a + V + NP [O]	NP + NP-ukku + Adv.+ Vt
	Usha foolishly answered the question.	uSaa keeLvikku muTTaaLtanamaaka viTaiyaLittaaL
6a	NP + V (AV) + NP [O] + Adv 12 (ex. Badly and well)	NP + NP-ukku + Adv. + V
	Kannan paid her well.	kaNNan avaLukku ndanRaaka campaLam koTuttaan
	Uma treated him badly	umaa avanai moocamaaka ndaTattinaaL
6b	2. NP + BE + adv12 + V (PV)	NP + NP-ukku + Adv. V
	She was well paid.	avaLukku ndanRaaka campaLam koTukkappaTTatu
	He was badly treated	avan moocamaaka ndaTattappaTTaan
6c	Adv (somehow) + …V….	NP + Adv + NP-ai + Vt
	Somehow they did it.	avarkaL eppaTiyoo atai ceytaarkaL
	… V. Adv (somehow)	NP + NP-ai + Adv + Vt
	They did it somehow.	avarkaL atai eppaTiyoo ceytaarkaL

Sno	**Adverbial Patterns of English**	**Parallel Patterns in Tamil**
Adverb of Place		
1a	NP + V + Adv2 (away, everywhere, here, nowhere, somewhere, there etc.,)	NP + Adv + V
	Usha waits outside	uSaa veLiyee kaattirundtaaL

	Raja sent her aboard	raajaa avaLai veLindaaTRiRku anuppinaan
	Write it there.	itai angkee ezhutu
1b	NP + V + PP + Adv2	NP + NP-ai + Adv + V
	Kannan looked for it evrerywhere	kaNNan atai ellaa iTangkaLilum teeTinaan
1c	NP + Vt + NP + Adv2	NP + NP-ai + Adv + V
	Keep the book somewhere	puttakattai engkeeyuaavatu vai
2a	NP + Vt + (NP) + Adv21 (ex. Somewhere and anywhere)	NP + (NP-ai) + Adv + V
	Usha has seen it somewhere	uSaa atai engkoo paarttirukkiRaaL
	Uma hasn't gone anywhere	umaa atai engkum paarttirukkavillai
	Kannan has gone somewhere	kaNNan engkoo pooyirukkiRaan
	Kannan hasn't gone anywhere	kaNNan engkum pooyirukkavillai
2c	Here / there + BE/COME/GO + NP [S]	itoo/atoo + NP atooatto + NP + V
	Here's Usha's friend.	itoo uSaavin ndaNpar.
	There goes my wife.	itoo en manaivi pookiRRaaL.
	Here comes Raja.	itoo raajaa varukiRaan.
	There comes the elephant.	atoo yaanai varukiRatu.
2b	There/here + NP [=Personal pronoun] +V	atoo + NP + V
	There he goes	atoo avan pookiRaan
	Here he comes	atoo avan varukiRaan

Sno	**Adverbial Patterns of English**	**Parallel Patterns in Tamil**
Adverb of Time		
1a	Adv 31 (Afterwards, eventually, lately, now, recently, soon, etc.,) +NP + V + NP [IO] + NP [O]	Adv + NP + NP-iTam + NP-ai + V
	Eventually Usha told Uma the secret	muTivaaka uSaa umaaviTam irakaciyattaic connaaL
	Usha told Uma the secret eventually	uSaa umaaviTam irakaciyattaic connaaL muTivaaka.

1b	NP + V + Adv.32 (before, early, immediately, and late)	NP + Adv + V
	Kala came early.	Kalaa munnareettil vandtaaL
2	NP + V + Adv 32	NP + Adv. + V
	Uma has gone there before	Umaa munnareetil angku pooyviTTaaL
	Let's start late	ndaam taamatamaaka puRappaTalaamaa
	Come immediately	viraivaaka vaa.
3	NP + V-Perf + Adv33 (since and ever since)	Adv. + NP + NP-ai + V
	Uma left Mysore in 1998	Umaa 1998-il maicuurukkuc cenRaaL atanPiRatu
	I haven't seen her since	ndaan avaLai Paarkkavilleii
4a	1. NP + V + (NP [O]) + Adv.34 (Yet and still)	NP +(NP-ai) + Adv + V
	Usha hasn't come yet.	uSaa ituvarai varavillai.
	Uma hasn't seen him yet	umaa avanai ituvaraip paarkkavillai
4b	NP + Adv34 + Vt + NP [O]	NP + Adv + (NP-ai) + V
	Kalaa hasn't yet finished the work I gave her a week ago.	kalaa ituvarai ndaan oruvaarattiRku munnar koTutta vellaiyai muTikkavillai

Adverbs of frequency 1.NP + V + Adv4 (Always, continually, frequently, occasionally, often, usually, once, etc.,) + Adj/NP NP + Adv + V Uma always comes late 2. NP + Adv4 + V Jaya often comes late. Kala is often late Kannan seldom visits Uma	NP + Adv + V uSaa vazhakkamaaka usha is usually happy. makizhcciyuTan irukkiRaaL NP + Adv + V Umaa epootum taamatamaaka varukiRaaL jayaa epootum taamatamaakka varukiRaaL kalaa aTikkaTi taamatamaaka varukiRaaL kaNNan eppootaavatu umaavai paarkka vuruvaan
1. NP + Aux1 + Adv 4 + Aux 2 + V Uma has often been warned Has Uma ever been warned?	NP + Adv + V umaa epootum eccarikkap paTukiRRaaL

2. Aux 1 + NP + Adv1 + V	NP + Adv + V umaa epootaavatu eccarikkap paTTaaLaa?
Adv 1 + Adv 41 + NP + V Secretly ever did Uma try to meet Usha? Adv 42 + NP + V Seldom have Usha beard such a speech	Adv + Adv + NP + NP –ai V irakaciyamaaka epootaavatu umaa uSaavai candtikka muyanRaaLaa? Adv + NP + NP-ai + V aritaakat taan uSaa appaTippaTTa peecai keeTTirukkiRaaL
There + BE + Adv4 (hardly, scarcely, and barely) + NP There is hardly any money left. NP + Adv 4 + V	NP + Adv + V paNam konjam kuuTa miitamillai NP + Adv + V
Usha hardly ever visit her friends	uSaa aritaakattaan taan ndaNparkaLai candtippaaL
Sentence adverbs: These modify the whole sentence / clause and normally express the speaker opinion. NP + BE + Adv 5 (actually, apparently, certainly, definitely, perhaps, surely, etc.,) +Adj. Usha is certainly right Uma is apparently happy	NP + Adv + Adj + BE uSaa ndiccayamaaka cari umaa veLippaTaiyaaka makizhcciyaaka irukmkiRaaL
NP + Adv5 + V … Kannan definitely looks happy	NP + Adv + ….V kaNNan ndiccayamaaka makizhcciyaakat terikiRaan
NP + Aux 1 + Adv5 + Aux2 + V Uma would obviously have gone NP + Aux + Adv5 + V Usha will surely come	NP + Adv + Aux1 + V umaa ndiccayamaakp Pooy iruppaaL NP + Adv + V uSaa kaTTaayamaaka varuvaaL
Adv 5 + NP + V…. Apparently Uma looks happy. NP + V … + Adv5 Uma looks happy apparently….	Adv + NP + V veLippaTaiyaaka umaa makizhcciyaakat terikiRaaL umaa makizhicciyaakat terikiRaal veLippa Taiyaaka
NP + V…. + Adv 51 (definitely) Uma will like Usha definitely NP + V…. + Adv52 (perhaps and possibly) Perhaps Uma will like Usha	NP + Adv + NP-ai + V umaa ndiccayamaaka uSaavai virumpuvaaL Adv + NP + (NP-ai + V oruveeLai umaa uSaavai virumpuvaaL
Adv 53 (admittedly, frankly, honestly, etc), NP + V…. Honestly, Usha has won the first prize.	Adv + NP + (NP) + V unmaiyaaka uSaa mutal paricu peRRirukkiRaaL

Adverbs of degree NP + BE Adv6 (quite, almost, barely, completely, enough, quite, rather, etc.,) + Adj Usha is quite happy. Uma is extremely beautiful.	NP + Adv + Adj + iru usaa mikavum makizheciyaaka irukkiRaaL uSaa mikavum azhakaaka irukkiRaaL NP + Adv + … + V
NP + BE + Adv 6 (quite, almost, barely, completely, enough, quite, rather, etc.) +V … Kalaa was completely covered with mud NP + HAVE + Adv6 + V….	uSaa makizhcciyaaka irukkiRaal umaa mika mika azhakaaka iRukkiRaaL kalaa muzhuvatum tozhiyaal muuTappaTTaL
Bava had almost reached Chennai. NP + BE + Adj + Adv61 (enough) The knife isn't sharp enough That food is not good enough	NP + Adv + …. + V Bavaa kiTTattaTTa cennaiyai aTaindtu viTTaaL NP + NP + BE Kattikku kuurmai pootaatu NP + Adj + BE caappaaTu ndanRaaka illai
NP + Adv 62 + V Uma almost fell down	NP + Adv + V umaa kiTTattaTTa vizhundtuviTTaaL
NP + V + Adv 63 (only) Usha ate only banana Kala only gave me her pen. NP + V + NP + PP + Adv 63 (only) Kala gave her pen to me only.	NP + …. + Adv + V uSaa vaazhaippazham maTTum caappiTTaaL kalaa maTTum Taan peenaa koTuttaal NP + … + Adv + V Kala Peena koTuttaal enaakku maTTumTaan
NP + Aux + Adb 64 (Just) + V… Uma has just gone home Usha has just finished her work	NP + Adv + …V + Aux umaa ippootu taan viTTukkup pooyirukkiRaaL uSaa ippootutaan veelaiyai muTittaal.

4.3.5 Paralles in adpositional phrases

Adposition is cover term used to incorporate preposition, postposition and cases markers. For the sake of correlation we have to take into account the case suffixes of Tamil also under adposition. The difference, as we are well aware of, is that the case suffixes are inflectional elements of nouns and pronouns, whereas postpositions are loosely added after the oblique forms or case inflected forms of

nouns and pronouns. As both of them are used to express different case relations, they are not distinguished from one another for the sake of computation.

The following points are the out come of typological correlation of adpositional phrases in English and Tamil.

1. English generally makes use of prepositions to denote the case relation existing between Vern and noun phrase. But Tamil mostly makes use of case suffixes to denote various case relations. Of course, Tamil too makes use of postpositions at par with English prepositions.
2. Fairly obviously word order is an alternative to case marking in distinguishing subject from object in languages like English. In English the word order also distinguishes the patient object from the recipient or beneficiary object in double object constructions where the patient object always follows the other object:

    She gave me good marks.
    She cut me a bunch of dahlias.

3. It has frequently been observed that there is a correlation between the presence of case marking on noun phrases for the subject-object distinction and flexible word order and this would appear hold true for Tamil.
4. Typologically it appears that there is a tendency for languages that mark the subject-object distinction on noun phrases to have the basic order of subject-object-verb (SOV), and conversely a tendency for languages lacking such a distinction to have order subject-verb-object (SVO). This statement appears to hold true for English and Tamil.

The following table illustrates the transfer of prepositional phrases in English into Tamil.

Preposition + NP in English	**Prepositional Relations / meaning**	**NP + Postposition** **IN TAMIL**
At + NP He is standing at the bus stop.	Dimension type o position	NP-oblique+ - *il* Avan peerundtu ndilaiyattil ndiRkiRaan

To + NP He went to Chennai Give it to me.	Dimension type 0 Destination	NP-oblique + *ku/itam* Avan cennaikkuc cenRaan NP-oblique + *iTam* enniTam koTu
On + NP The book is on the table	Dimension type ½ position (line or surface)	NP-oblique +*il/meel* meecai il/meel puttakam irukkiRatu
On (to) + NP He fell on (to) the floor.	Dimension type ½ destination (line or surface)	NP-oblique+*il* avan taraiyil vizhundtaan
In He is in the village. In (to).	Dimension type 2/3 position (area or volume)	NP-oblique+*il* Avan kiraammattil irukkiRaan.
Kannan dived in (to) the water.	Dimension type 2/3 destination (area or volume)	NP-oblique + *il* kaNNan taNNiiril kutittaan
Away from (=not at) + NP He is away from Chennai	Dimension type o position	NP – oblique + *il + illai* avan cennaiyil illai
Away form + NP He went away form Chennai	Dimenstion type 0 destination	NP-oblique +accusative + *viTTu* svan cennaiayai viTTup poonaan
Off + NP The books were off (=not on) the shelves. Off + NP He took the book off the shelves Out of (=not in)+NP He is out of the office.	Dimension type ½ position (line or surface) Dimension type ½ destination (line or surface) Dimension type 2/3 position (area or volume)	NP-oblique *+-il + illai* puttakangkaL SelpukaLil illai NP-oblique + - *il + irundtu* avan SelpukaLilirundtu Puttakattai eTuttaan Avan aluvalakattil illai NP-oblique *+-il + illai*
Out of + NP He went out of the office	Dimension type 2/3 destination (area or volume)	NP-oblique+*-il + irundtu + veLiyee* Avan aluvalakattilirundtu veLiyee poonaan
Above / over/ On top of + NP The lamp is hanging over the head.	SUPERIOR	NP-oblique + (dative) + *meel/meelee* talaikku meelee viLakkut tongkukiRatu
Below / under/ underneath/ beneath+ NP The dog is lying under the table. INTERIOR in front of +{ NP The house is in	INFERIOR ANTERIOR	1. Inferior location ‘under’ NP-oblique+*in+ kiizh/kiizhee/aTiyil* 2. Inferior location ‘below’ NP-oblique + dative + *kiizh/kiizhee* meejaiyin aTiyil ndaay kiTakkiRatu NP-oblique + dative + *munnaal, munp/mun/munnar/mundti* Koovilukku munnaal viiTu

front of the temple. Behind + NP The house behind the temple Into / inside +NP He is inside the house Out of / outside + NP He went out of the house.	 POSTERIOR INTERIOR EXTERIOR	irukkiRatu NP-oblique + dative + pinnaal, pin, pinup, pinnar, pindti Koovilukkup pinnaal viiTu irukkiRatu. NP-oblique+dative + *uL, uLLee* Avan viiTTukkuL irukkiRaan NP-oblique + dative + *veLiyee* Avan viiTTuku veLiyee cenRaan.
Near / by / beside / by the side of / at the side of + NP He went near her	NEAR	NP -oblique+ dative + *arukil* Pakkattil/KiTTee Avan avaL pakkattil cenRaan.
With + NP The onion is lying with potato.	In the same place as position	NP-oblique + *ooTul uTan* venkaayam uruLaikkizhankuTan kiTakkiRatu
With + NP He wants with her.	In the same place as Accompainment	NP-oblique + *ooTu / uTan / kuuTee* avan avaLooTu cenRaan
Betweenm, amid, amidst, among, Among + NP The Minister stood among the People.	BETWEEN	NP + dative + *iTaiyil/ ndaTuvil* mandtiri makkaLukku ndaTuvil ndiRkiRaar
Beyond + NP The school is beyond the temple.	ULTERIOR	NP-oblique + accusative + *taaNTi* NP-oblique + dative + appaal koovilukku appaal paLLikkuuTam irukkiRatu
Opposite to + NP The house is opposite to the temple	 CITERIOR	NP + dative + etiree / etiril / etirkku/etirttaaR poola Koovilukku etiree viiTu irukkiRatu
Around +NP The trees are around the house.	CITERIOR CIRCUMFE–RENTIAL	NP-oblique + accusative + cuRRi ViiTTaic cuRRi marangkaL irukkinRana.
Across + NP They went across the river.	ACROSS	NP-oblique + accusative + *taaNTi* avarkaL aaRRait taaNTi cenRaarkaL. NP-oblique-in + *kuRukee* avarkaL aaRRin kuRukkee cenRaarkaL
Through + NP He went through	THROUGH	NP + *vazhiyaaka* avan kaaTTu vazhiyaakaP poonaan.

The forest.		
Along + NP He went along the road.	ALONG	NP + *vazhiyaaka* avan caalai vazhiyaakap poonaan.
Towards + NP He went towards The park	TOWARDS	NP-oblique + accusative + *ndookki. paarttu* avan puungkaavai ndookkic cenRaan.
From + NP He went from house.	SOURCE	NP-oblique + locative *il* + *irundtu*, NP + accusative + *viTTu* avan viiTTai viTTuc cenRaan.
To + NP He gave her money.	GOAL	NP-oblique + *ku/ iTam* avan avaLiTam paNam koTuttaan.
Because of + NP He came there because of her.	CAUSE	NP-oblique + *aal* avan avaLLal angku vandaan NP + kaaraNamaaka avan avaL kaaraNamaaka vandtaan.
For + NP He came there for seeing her.	PURPOSE	NP-oblique + dative + *aaka* avan avaLaip paarpatarkaaka angku vandaan
With + NP He treated her with respect.	MANNER	NP – oblique + *ooTu /uTan* NP + *aaka* avan avaLai mariyaataiyooTu ndaTattinaan.
By means of/ by + NP..I came by bus	MEANS	NP + oblique + aal ndaan pusil / pusaal vandteen
By + NP He beat the animal by a cane.	INSTRUMENT	NP-oblique + *aal* avan andta vilangkai pirampaal aTittaan. NP + accusative + vaittu / koNTu avan andta vilangkai pirampaal koNTu aTitaan.
About + NP He talked about her.	ABOUT	NP-oblique + accusative + paRRi/ kuRittu avan avaLaip paRRi peecinaan
In connection with + NP He went to Chennai in connection with his busniness	CONNECTION	NP-oblique + accusative + oTTi avan tan viyaapaarattai oTTi cennai cenRaan
For + NP He struggled for her.	SUPPORT	NP-oblique + dative + aaka, veeNTi avan avaLukku veeNTi pooraa Tinnaan.
Against + NP He fought against them.	OPPOSITION	NP-oblique + dative + *etiraaka* avan avarkaLukku etiraaka caNTaiyiTTaan
Except for / with the exception of / excepting / except /	EXCEPTION	NP + accusative + tavira tavittuozhiya usaavait tavira elloorum

but for/barring + NP All except Usha came to office.		aluvalakattiRku vandaarkaL
Istead of + NP He drank coffee Instead of tea.	SUBSTITUTION	NP + dative + *patilaaka* avan teendiirukkup patilaaka kaappi arundinaan
But for + NP But for him I have not bought the gift.	NEGATIVE CONDITION	NP + *illaaviTTaal/allaamal* avan illaaviTTaal ndaan inda paricai vaangki irukka maaTTeen.
With / out of - NP He walks with walking stick	INCREDIENT	NP-oblique + *aal* NP-oblique + (ai) + *koNTu/vaittu* avan ndaTakool koNTu ndaTandtaan

4.3.6 Paralles in Phrasal Co-Ordination

There are different types of phrasal co-ordination

Type of coordination	In English	In Tamil
Coordination of noun phrases	1. NP and NP Noun phrases are commonly conjoined Ram and Prem are brothers. 2. NP or NP Ram or Prem will come 3. Either NP or NP Either Ram or Prem did it. 4. Neither NP nor NP Neither Ram nor Prem did it.	1. NP – *um* NP - *um* cakootarakaL raamum pireemum raamoo pireemooi varuvaarkaL 2. NP – *oo* NP – *oo* 3. NP – *oo* NP – *oo* V – negative ramoo pireemoo atai ceyyavillai. (Note in Tamil the clitic oo can be replaced by negative element *allatu*.) 4.raam allatu pireem iruvarumee ceyyavillai
Coordination of more than two noun phrases	NP, NP… and / or NP *And* and *or* can link more than two NPs, and all but the final instance of the conjunctions can be omitted. We congratulated Ram, Prem, and Beem, Det and / or Det	1. Conjunction NP – um, NP– *um*,NP-um ndaangkaL raamaiyum, preemaiyum, piimaiyum paaraaTTinoom. 2. Disjunction NP – oo, NP – *oo* ndaangkaL raamaiyoo, preemaiyoo pimaiyoo paraaTTinoom.
Coordination of determiners	Demonstrative can be linked to each other or to other determiners in the NP. Take this and that Take this (pen) and that pen.	1. Det-*um* Det-*um* itaiyum ataiyum eTu 2. Det + N + um + Det + N + um Indta peenaavai – um andta peenaavai – um eTu. Indta peenavaiy – oo andta peenavai – *oo* eTu.
Coordination of	Adj P and / or Adj P	AdjP + um / *oo* + AdjP+ um/*oo*

adjectival phrases	Adjectives both predicative and attributive can be conjoined. She is beautiful and smart.	avaL azhakaakavum keTTikaarattanamaakavum iRukkiRaaL
Coordination of adverbial phrases	Adv Pnd/or Adv P Adverbials and dependent clause can be conjoined. I can announce it loudly or by using a speaker.	Adv+*um/oo* + Adv + *um/oo* ennall itai captamaakavoo allatu olipperukki koNToo aRivikka iyalum
Coordination of prepositional phrases	PP and / or PP Prepositional phrases can also be conjoined. He looks for his pen inside and outside the box The test in April and in May is postponed.	PP + um/oo + PP + um/oo avan tan peenaavai peTTikku uLLeeyum veLiyeeyum teeTinaan NP+maRRum+NP eeppiral maRRum meey teervukaL ottivaikkap paTTirukkinRana.

4.4 Summary

The parallel structures in English and Tami at the sentential level, clause level, and pharase level have been extracted from English Tamil parallel corpoara. The extracted parallel structures reveal the correlating syntactic structures of the two languages. The correlative study tries to explore the commonalities and differences in the structure of English and Tamil from the point of view of computation to build machine translation system using parallel corpus to translate English into Tamil. It has been noticed that the two language deviate from one another from the point of view of English as language of SVO and Tamil as language of SOV, i.e., verb final language. While English makes use of prepositions to link nominal arguments with verbs, Tamil makes use of postpositions and case markers to serve the same purpose. The absence of regular case inflections in the case of English makes it rigid in its word order and the presence of case inflections in Tamil makes it more flexible in its word order. English distinguishes subject form object by means of the position, i.e., word order, where as Tamil does it by case inflections. Relative clause in English is after the head noun, which is attributed and in Tamil it comes before the head noun. The infinitive clause in English comes after the main clause, whereas in Tamil it is comes before the main clause. That-clause complement occurs at the right side of the main clause in English, whereas it occurs at the left side of the main clause in Tamil. Interrogation is effected by changing the order of the words, i.e., by

moving an auxiliary verb to the initial position before subject. In Tamil interrogation is effected by suffixing interrogative clitic or by making use of interrogative pronouns. In English, the auxiliary verbs and the interrogative words occur in the initial position of the construction. In Tamil, the interrogative particles occur in the final position of any word in the construction. All these correlative features have to be taken into account while preparing the parallel corpus for Englihs-Tamil machine transltion based on statical approach.

Chapter 5
English to Tamil Machine Translation System
By using parallel Corpus

5. Introduction

The rule based approach dominated the area of the machine translation until 1989, when IBM introduced the Statistical Machine Translation approach inspired by Weaver memorandum of 1949, the availability of parallel corpus in the Canadian parliament and the advantages of empirical approach over the rule based approach. During the years of 1993 – 1999, there are only a few activities related to statistical machine translation due to the lack of open source tools for statistical machine translation. Later, when JHU workshop implemented open source tools for statistical machine translation tools for IBM statistical machine translation model in 1999, the research in statistical machine translation approach has started dominating Natural Language Processing till now.

5.1 On the subject of SMT

Statistical Machine Translation (SMT) is a data oriented statistical framework for translating text from one natural language to another, rooted in the knowledge extracted from bilingual corpus. Unlike rule based MT systems, this approach does not require any language specific linguistic knowledge to perform the translation. The only requirement for the statistical machine translation system is a huge parallel

corpus. Performance of the statistical machine translation system is largely driven by the availability of the sentence aligned bilingual corpus. SMT research gained momentum in early 1990's after the availability of Hansar Canadian parliamentary proceedings (in English and French) in digital format. Many algorithms were developed to identify the sentence pairs automatically from the bilingual corpus. Brown *et al.* (1993) proposed a series of statistical models known as IBM translation models which became the basis for word-based statistical machine translation systems.

5.1.1. Statistical Machine Translation and the Noisy Channel Model

Statistical Machine Translation is founded upon the assumptions of the *Noisy Channel Model* and *Bayes Rule* which help 'decompose' the complex probabilistic model that needs to be built for estimating the probability of a sentence in a source language (f) being translated into a particular target language sentence (e). Using the notation common in the literature this decomposition can be stated as:

$$P(e|f)=P(e)*P(f|e)/P(f)$$

Since predicting in a statistical model corresponds to identifying the most likely translation, maximizing the above over all possible target sentences (e) gives the estimation:

$$\text{argmax}_e P(e|f)=\text{argmax}_e P(e) *P(f|e)$$

The main benefit gained by the above decomposition is that the burden of accuracy is moved away from the single probability distribution P(e|f) to two independent probabilities P(e) and P(f|e). The former is known as the 'language model' (for language e) while the latter is known as the 'translation model' (for predicting source sentences, f , from target sentences e). While it would be impossible to estimate such a language model, the literature on using n-gram (mainly bi-gram and tri-gram) models for estimating sentence probabilities of a given language have matured over the past two decades. The estimation of the translation model would not be too difficult if machine readable dictionaries with frequency statistics were available. While this is impractical for even the most well studied

languages, the dependence of such counts on the genre of the texts under consideration make it less than optimal.

This is where work carried out by Brown et. al. at IBM stepped into providing a bootstrapping model building process. Beginning with the very simple word-for-word translation lexicon building models (IBM Models 1 and 2), this process constructs ever more sophisticated Models (3, 4 and 5) which account for more and more flexibility in the underlying assumptions (e.g. a single word in the source language may be translated by more than a single target word, and may appear in another part of the sentence). Intuitively, once the translation model performs its task of predicting a set of possible (good and bad) candidate translations for a particular source sentence, the (target) language model will calculate the probability of such sentences being acceptable in the language in order to select the best translation. It is this 'sharing of the burden of accuracy' between the two models that has been at the heart of the relative success of the SMT approach.

5.1.2 Advantages of SMT

The most frequently cited benefits of statistical machine translation over traditional paradigms are:

- **Better use of resources**

    1. There is a great deal of natural language in machine-readable format.
    2. Generally, SMT systems are not tailored to any specific pair of languages.
    3. Rule-based translation systems require the manual development of linguistic rules, which can be costly, and which often do not generalize to other languages.

- **More natural translations**
    1. Rule-based translation systems are likely to result in Literal translation. While it appears that SMT should avoid this problem and result in natural translations, this is negated by the fact that using statistical matching to translate rather than a dictionary/grammar rules approach

can often result in text that include apparently nonsensical and obvious errors.

5.1.3 Challenges with statistical machine translation

Problems that statistical machine translations have to deal with include:

- **Sentence alignment**

In parallel corpora single sentences in one language can be found translated into several sentences in the other and vice versa. Sentence aligning can be performed through the Gale-Church alignment algorithm.

- **Compound words**
- **Idioms**

Depending on the corpora used, idioms may not translate "idiomatically". For example, using Canadian Hansard as the bilingual corpus, "hear" may almost invariably be translated to "Bravo!" since in Parliament "Hear, Hear!" becomes "Bravo!".

- **Morphology**
- **Different word orders**

Word order in languages differs. Some classification can be done by naming the typical order of subject (S), verb (V) and object (O) in a sentence and one can talk, for instance, of SVO or VSO languages. There are also additional differences in word orders, for instance, where modifiers for nouns are located, or where the same words are used as a question or a statement.

In speech recognition, the speech signal and the corresponding textual representation can be mapped to each other in blocks in order. This is not always the case with the same text in two languages. For SMT, the machine translator can only manage small sequences of

words, and word order has to be thought of by the program designer. Attempts at solutions have included re-ordering models, where a distribution of location changes for each item of translation is guessed from aligned bi-text. Different location changes can be ranked with the help of the language model and the best can be selected.

- **Syntax**
- **Out of vocabulary (OOV) words**

SMT systems store different word forms as separate symbols without any relation to each other and word forms or phrases that were not in the training data cannot be translated. This might be because of the lack of training data, changes in the human domain where the system is used, or differences in morphology.

5.2 The Components of Statistical Machine Translation

The SMT system is based on the view that every sentence in a language has a possible translation in another language. A sentence can be translated from one language to another in many possible ways. Statistical translation approaches take the view that every sentence in the target language is a possible translation of the input sentences. Figure 5.1 gives the outline of Statistical Machine Translation system.

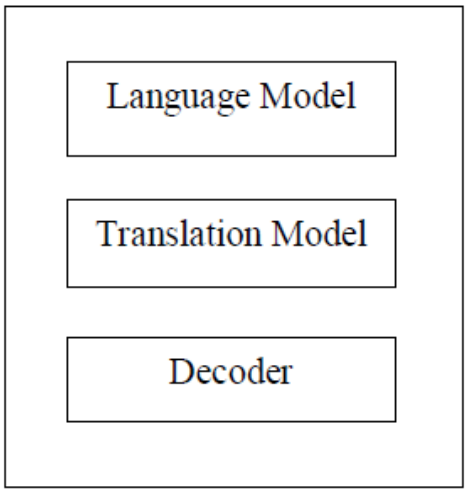


Figure 5.1 Outline Statistical Machine Translation system

5.2.1 Language Model

A language model gives the probability of a sentence. The probability is computed using *n-gram* model. Language Model can be considered as computation of the probability of single word given all of the words that precede it in a sentence. The goal of Statistical Machine Translation is to estimate the probability (likelihood) of a sentence. A sentence is decomposed into the product of conditional probability. By using chain rule, this is made possible as shown in 5.1. The probability of sentence P (S), is broken down as the probability of individual words P(w).

$$P(s) = P(w1, w2, w3,....., wn)$$
$$=P(w1)\ P(w2|w1)\ P(w3,|w1w2)\ P\ (w4|w1w2w3)\ldots P(wn|w1w2\ldots wn\text{-}1)) \ldots (5.1)$$

In order to calculate sentence probability, it is required to calculate the probability of a word, given the sequence of word preceding it. An *n-gram* model simplifies the task by approximating the probability of a word given all the previous words.

An *n*-gram of size 1 is referred to as a *unigram*; size 2 is a *bigram* (or, less commonly, a *digram*); size 3 is a *trigram*; size 4 is a *four-gram* and size 5 or more is simply called a *n-gram*.

Consider the following training set of data given in Figure 5.2:

There was a King
He was a strong King.
King ruled most parts of the world.

Figure 5.2. Training set of data for LM

Probabilities for bigram model are as shown below:

$$P(there/<s>) = 0.67\ P(was/there) = 0.4\ P(king/a) = 1.0\ P(a/<s>)=0.30 \quad \ldots(5.2)$$

P(was/he) = 1.0 P(a/was) = 0.5 P(strong/a) = 0.2 P(king/strong)=0.23 ... (5.3)
P(ruled/he) = 1.0 P(most/rules) = 1.0 P(the/of) = 1.0 ...(5.4)
P(world/the) =0.30 P(ruled|king)=0.30 ... (5.5)

The probability of a sentence: 'A strong king ruled the world', can be computed as follows:

P(a/<s>)*P(strong/a)*
P(king|strong)*P(ruled|king)*P(the/ruled)*P(world|the)
=0.30*0.2*0.23*0.30*0.28*.0.30
=0.00071 ... (5.6)

5.2.2 Translation Model

The role of the translation model is to find $P(f \mid e)$ the probability of the source sentence f given the translated sentence e. Note that it is $P(f \mid e)$) that is computed by the translation model and not $P(e \mid f)$The training corpus for the translation model is a sentence-aligned parallel corpus of the languages F and E.

It is obvious that we cannot compute $P(f \mid e)$ from counts of the sentences *f* and *e* in the parallel corpus. Again, the problem is that of data sparsity. The solution that is immediately apparent is to find (or approximate) the sentence translation probability using the translation probabilities of the words in the sentences. The word translation probabilities in turn can be found from the parallel corpus. There is, however, a problem - the parallel corpus gives us only the sentence alignments; it does not tell us how the words in the sentences are aligned.

A word alignment between sentences tells us exactly how each word in sentence *f* is translated in *e*. The problem is getting the word alignment probabilities given a training corpus that is only sentence aligned. This problem is solved by using the Expectation-Maximization (EM) algorithm.

5.2.2.1 Expectation Maximization

The key intuition behind EM is that if we know the number of times a word aligns with another in the corpus, we can calculate the word translation probabilities easily. Conversely, if we know the word translation probabilities, it should be possible to find the probability of various alignments. However, if we start with some uniform word translation probabilities and calculate alignment probabilities, and then use these alignment probabilities to get better translation probabilities, and keep on doing this, we should converge on some good values. This iterative procedure, which is called the Expectation-Maximization algorithm, works because words that are actually translations of each other co-occur in the sentence-aligned corpus.

5.2.2.2 Different Translation Models

As explicitly introduced by IBM formulation as a model parameter, word alignment becomes a function from source positions j to target positions i, so that $a(j) = i$. This definition implies that resultant alignment solutions will never contain many-to-many links, but only many-to- one , as only one function result is possible for a given source position j.

Although this limitation does not account for many real-life alignment relationships, in principle IBM models can solve this by estimating the probability of generating the source empty word, which can translate into non-empty target words. However, as we will see in the following section, many current statistical machine translation systems do not use IBM model parameters in their training schemes, but only the most probable alignment (using a Viterbi search) given the estimated IBM models. Therefore, in order to obtain many-to-many word alignments, usually alignments from source-to-target and target-to-source are performed, and symmetrization strategies have to be applied.

5.2.2.2.1 Word-based Translation Model

In word-based translation model, translation elements are words. Typically, the number of words in translated sentences is different due to compound words, morphology and idioms. The ratio of the length of sequences of translated words is called fertility, which tells how many English words, each native word produces.

Simple word-based translation is not able to translate language pairs with fertility rates different from one. To make word-based translation systems manage, for instance, high fertility rates, and the system could be able to map a single word to multiple words, but not vice versa. For instance, if we are translating from English to Tamil, each word in Tamil could produce zero or more English words. But there's no way to group two Tamil words producing a single English word.

An example of a word-based translation system is the, freely available, GIZA++ package, which includes the training program for IBM models and HMM models. The word-based translation is not widely used today comparing to phrase-based systems, whereas, most phrase based system are still using GIZA++ to align the corpus. The alignments are then used to extract phrase or induce syntactical rules. And the word alignment problem is still actively discussed in the community. Because of the importance of GIZA++, there are now several distributed implementations of GIZA++ available online.

Statistical machine translation is based on the assumption that every sentence t in a target language is a possible translation of a given sentence ‗e' in a source language. The main difference between two possible translations of a given sentence is a probability assigned to each, which is to be learned from a bilingual text corpus. The first statistical machine translation models applied these probabilities to words, therefore considering words to be the translation units of the process.

5.2.2.2.2 Phrase-based Translation Model

In phrase-based translation model, the aim is to reduce the restrictions of word-based translation by translating whole sequences of words, where the lengths may differ. The sequences of words are called blocks or phrases, but typically are not linguistic phrases but phrases found using statistical methods from corpora.

The job of the translation model, given a Tamil sentence T and an English sentence E, is to assign a probability that T generates E. While we can estimate these probabilities by thinking about how each individual word is translated. Modern

statistical machine translation is based on the intuition that a better way to compute these probabilities is by considering the behavior of phrases. The intuition of phrase-based statistical machine translation is to use phrases i.e., sequences of words as well as single words as the fundamental units of translation.

The generative story of phrase based translation has three steps. First, we group the source word into phrases $E_1, E_2, \ldots E_l$. Second, we translate each E_i to T_i. Finally, each phrase in the source is reordered.

The probability model for phrase based translation relies on a translation probability and distortion probability. The factor $\varphi(Ti \mid Ei)$ is the translation probability of generating source phrase T_i from target phrase E_i. The reordering of the source phrase is done by distortion probability d. The distortion probability in phrase based translation means the probability of two consecutive Tamil phrases being separated in English by a span of English word of a particular length. The distortion is parameterized by 1 d $(a_i - b_{i-1})$ where a_i is the start position of the source English phrase generated by the i^{th} Tamil phrase, and b_{i-1} is the end position of the source English phrase generated by $i-1^{th}$ Tamil phrase. We can use a very simple distortion probability which penalizes large distortions by giving lower and lower probability for larger distortion. The final translation model for phrase based machine translation is based on the equation (5.7).

$$P(T \mid E) = \prod_i \varphi(T_i \mid E_i)\ d(a_i - b_{i-1}) \qquad \ldots. \ 5.7$$

Phrase based models works in a successful manner only if the source and the target language have almost same in word order. Difference in the order of words in phrase based models is handled by calculating distortion probabilities. Reordering is done by the phrase based models. It has been shown that restricting the phrases to linguistic phrases decreases the quality of translation. By the turn of the century it became clear that in many cases specifying translation models at the level of words turned out to be inappropriate, as much local context seemed to be lost during

translation. Novel approaches needed to describe their models according to longer units, typically sequences of consecutive words or phrases.

The translation process takes three steps:

1. The sentence is first split into phrases - arbitrary contiguous sequences of words.
2. Each phrase is translated.
3. The translated phrases are permuted into their final order. The permutation problem and its solutions are identical to those in word-based translation.

Consider the following particular set of phrases for our example sentences:

Tamil	Netru	naAn	avaLai	pArththaen
English	yesterday	I	saw	her

Since each phrase follows are not directly in order, the distortions are not all 1, and the probability P (E | T) can be computed as:

$$P(E|T)=P(\text{yesterday}|\text{Netru})\times d(1) \times P(\text{i}|\text{naAn})\times d(1) \times P(\text{her}|\text{avaLai})\times d(2) \times P(\text{saw}|\text{pArththaen})\times d(2) \quad \text{….. 5.8}$$

Phrase-based models produce better translations than word-based models, and they are widely used. They successfully model many local re-orderings, and individual passages are often fluent. However, they cannot easily model long-distance reordering without invoking the expense of arbitrary permutation.

5.2.2.2.3 Factored Translation Model

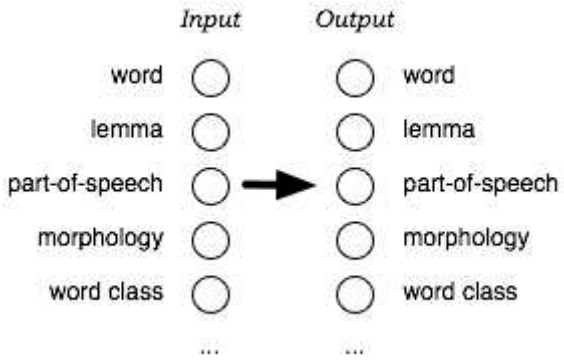
The current state-of-the-art approach to statistical machine translation, so-called phrase-based models, are limited to the mapping of small text chunks (phrases) without any explicit use of linguistic information, may it be morphological,

syntactic, or semantic. Such additional information has been demonstrated to be valuable by integrating it in pre-processing or post-processing.

However, a tighter integration of linguistic information into the translation model is desirable for two reasons:

- Translation models that operate on more general representations, such as lemmas instead of surface forms of words, can draw on richer statistics and overcome the data sparseness problems caused by limited training data.
- Many aspects of translation can be best explained on a morphological, syntactic, or semantic level. Having such information available to the translation model allows the direct modeling of these aspects. For instance: reordering at the sentence level is mostly driven by general syntactic principles, local agreement constraints show up in morphology, etc.

Therefore, we developed a framework for statistical translation models that tightly integrates additional information. Our framework is an extension of the phrase-based approach. It adds additional annotation at the word level. A word in our framework is not anymore only a token, but a vector of factors that represent different levels of annotation (see figure below).



5.2.3 The Statistical Machine Translation Decoder

The statistical machine translation decoder performs decoding which is the process of finding a target translated sentence for a source sentence using translation model and language model.

In general, decoding is a search problem that maximizes the translation and language model probability. Statistical machine translation decoders use best-first search based on heuristics. In other words, decoder is responsible for the search of best translation in the space of possible translations. Given a translation model and a language model, the decoder constructs the possible translations and look for the most probable one. There are a numerous decoders for statistical machine translation. A few of them is greedy decoders and beam search decoders. In greedy decoders, the initial hypothesis is a word to word translation which was refined iteratively using the hill climbing heuristics. Beam search decoders use a heuristic search algorithm that explores a graph by expanding the most promising node in a limited set.

5.3 Tools used for implementation of SMT System

Various tools are available for the development of Statistical Machine Translation. A SMT system for a pair of languages can be developed by using the combination of these tools. It shows some open source tools that are available to use. Freely downloadable Software is as follows:

- EGYPT system
  System from 1999 JHU workshop. Mainly of historical interest.
- GIZA++ and mkcls
  Franz Och. C++. GPL.
- Thot
  Phrase-based model building kit
- Phramer
  An Open-Source Java Statistical Phrase-Based MT Decoder
- Moses
  A new open-source phrase-based MT decoder with functionality beyond Pharaoh.
- Syntax Augmented Machine Translation via Chart Parsing
  Andreas Zollmann and Ashish Venugopal

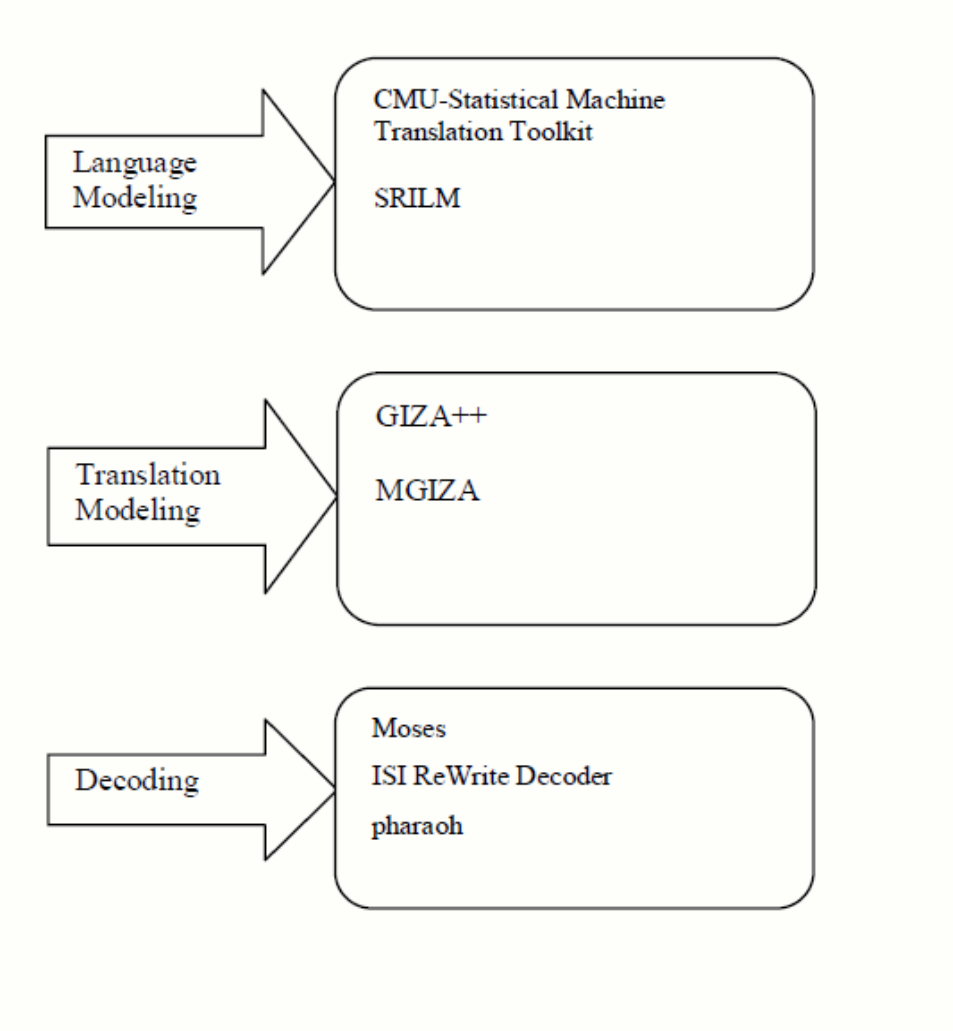


Figure 5.3: Statistical Machine Translation Tools

5.3.1 Language Model (LM) tools

There are many LM tools which are available. They are discussed as follows.

The CMU Statistical Language Modeling (SLM) Toolkit

The Carnegie Mellon University (CMU) Statistical Language Modeling Toolkit is a set of UNIX software tools designed to facilitate Language Modeling work for research purposes. It was written by Roni Rosenfeld, and released in 1994.

SRILM

SRILM is a toolkit for building and applying statistical Language Models (LMs) developed by SRI Speech Technology and Research Laboratory. It has been under development since 1995. SRILM is freely available for download.

5.3.2 Translation Model Tools

There are many TM tools which are available to be used for SMT systems. They are
discussed as follows.

GIZA++

GIZA++ is a tool developed by Franz Josef Och. and is an extension of GIZA developed by the Statistical Machine Translation team during the summer workshop in 1999 at the center for Language and Speech Processing at Johns-Hopkins University. This tool implements different models like HMM and also perform word alignment. GIZA++ is freely available for download.

MGIZA

MGIZA++ is a multi-threaded word alignment tool based on GIZA++. It extends GIZA++ in multiple ways. It provides the concept of multi-threading, and memory optimization. It can resume training from any stage, and continue training from any stage. MGIZA is freely available for download.

5.3.3 Decoder Tools

There are many different tools for the decoding stage of SMT system. They are discussed as follows.

Moses

Moses is a Statistical Machine Translation system developed by Hieu Hoang and Philipp Koehn at the University of Edinburgh that allows the automatic training of translation models for any language pair. All that is required is a collection of translated texts (parallel corpus). Moses works with SRILM to develop Language Model, and GIZA++ to develop Translation Model. Moses is freely available for download.

ISI ReWrite Decoder

ISI ReWrite Decoder is software that is used to perform decoding (searching) in development of Statistical Machine Translation systems. It works with CMUStatistical Language Modeling toolkit and GIZA++ to perform translations from Source Language to Target Language. It freely available for download and use at the link.

http://www.isi.edu/publications/licensed-sw/rewrite-decoder/

Pharaoh

Pharaoh is a Machine Translation decoder developed by Philipp Koehn as part of his PhD thesis at the University of Southern California and the Information Sciences Institute to aid research in Statistical Machine Translation. The decoder works with the SRI Language Modeling Toolkit. It can be obtained from link.

http://www.isi.edu/licensed-sw/pharaoh/

5.4 Existing Statistical MT Systems

There are following MT systems that have been developed for various natural language pair.

Google Translate

Google Translate is service provided by Google Inc. to translate a section of text, or a webpage, into another language. The service limits the number of paragraphs, or range of technical terms, that will be translated. Google translate is based on Statistical Machine Translation approach. It can translate text, documents, web pages *etc*.

Bing Translator

Bing Translator is a service provided by Microsoft, which was previously known as Live Search Translator and Windows Live Translator. It is based on Statistical Machine Translation approach. Four bilingual views are available:

- Side by side
- Top and bottom
- Original with hover translation
- Translation with hover original

5.5 Problem Statement

With each passing day the world is becoming a global village. There are hundred's of languages being spoken across the world. The official languages of different states and nations are also different according to their cultural and geographical differences.

5.5.1 Gap Analysis

Most of the content available in digital format is in English language. The content shown in English must be presented in a language which can be understood by the intended audience. There is large section of population at both national and state level who cannot comprehend English language. It has brought about language

barrier in the side lines of digital age. Machine Translation (MT), can overcome this barrier. In this thesis, a proposed Statistical Based Machine Translation system for translating English text to Tamil language has been proposed. English is the source language and the Tamil is the target language.

5.6 Development of Corpus

Statistical Machine Translation system makes use of a parallel corpus of source and target language pairs. This parallel corpus is necessary requirement before undertaking training in Statistical Machine Translation. The proposed system has used parallel corpus of English and Tamil sentences. A parallel corpus of more than 5000 sentences has been developed from which consist of small sentences and the life history of freedom fighters with reference to their trail in courts.

5.7 Architecture of English to Tamil Statistical Machine Translation System

The architecture forms the central role in making up SMT system. Language Model (LM), Translation Model (TM), decoder are used in undertaking SMT. Language Model is prepared from the target language. Decoder gives the probability of target sentence given the source sentences. The architecture of the system is shown in Figure 5.3.

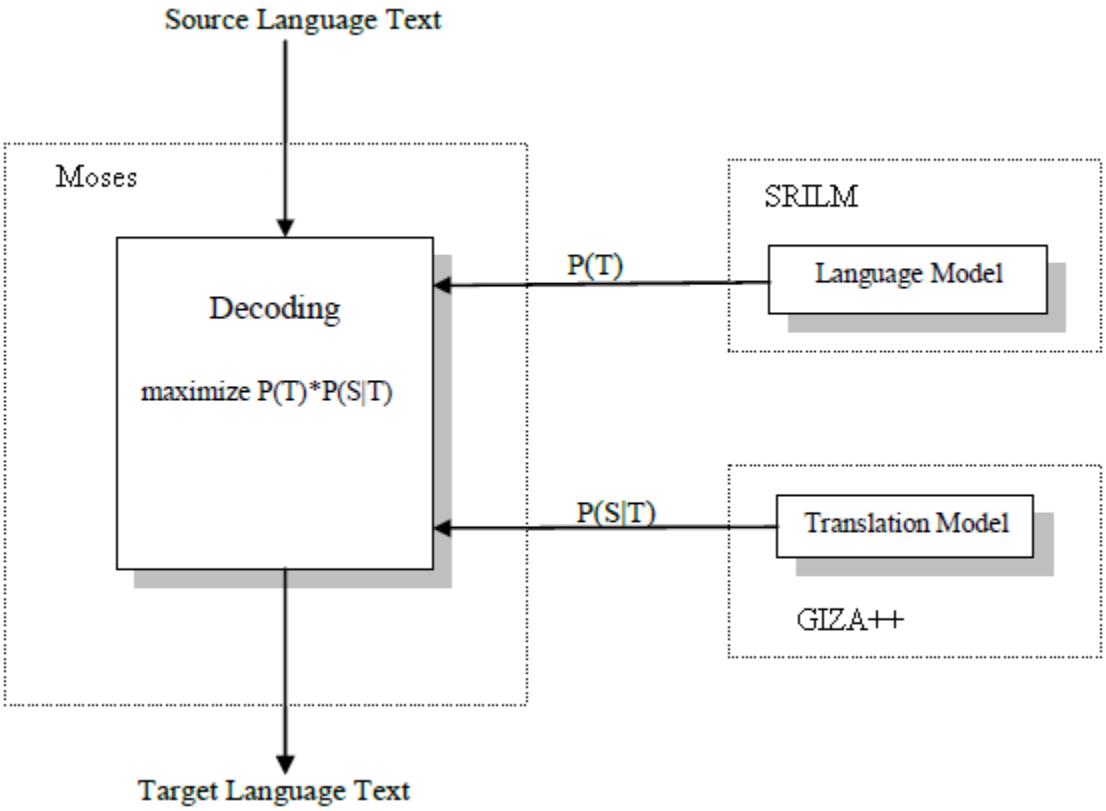


Figure 5.4: Architecture of Statistical Machine Translation system

5.7.1 Architecture for Language Model

Language Model (LM) gives the probability of a sentence. The probability of a sentence depends upon the probability of individual words. *n-gram* is a sequence of words. LM is developed for the target language. If '*T*' is the target language, LM computes '*P(|T)*' and feed this input to the decoder software. SR International's Language Model (SRILM) for LM is used. SRILM is available freely for research purposes from their website

http://www.speech.sri.com/projects/srilm/download.html.

5.7.2 Architecture for Translation Model

The Translation Model (TM) computes the probability of source sentence '*S*', for a given target sentence *'T'*. Mathematically, the probability being computed by TM is given as, *P(S|T)*. Translations can be done word based or phrase based. The

output of TM is fed into Moses decoder. *GIZA++* along with *mkcls* is used to develop Translation model, which is developed.

5.7.3 Architecture for Decoder

The decoder maximizes the probability of the generated sentence. It makes use of the a*rgmax ()* function to maximize the probability. *Moses* software which is freely available under open source licenses is used for decoder. *Moses* is compatible with SRILM and *GIZA++*. *Moses* decoder accepts as input the source language text and generates the target language text. The probability files are accepted from TM and LM. The decoder can be set in interactive mode to for doing translation.

5.8 Preparation of Data

Preparation of data involves tokenizing, cleaning, lowercasing the corpus. Before undertaking the training of the system the data must be pre-processed. The issues which need to be addressed in parallel corpus are as follows:

- To set the environment variable LC_ALL to C in Linux environment.
- The software needs one sentence per line. So there should be no empty lines in the corpus.
- The sentences having word limit more than 40 words are removed. The sentences having word limit from 1-40 are not removed.
- All sentences of parallel corpus need to be in lowercased. The uppercased sentences need to be changed to lower case.

For the preparation of data, used in proposed system, PERL scripts have been used.

5.8.1 Tokenizing the corpus

Tokenizing of corpus makes use of a Perl script. The input to this script is the raw corpus and the output is tokenized corpus. The script executed as given in 5.9.

zcat corpus_new4.en.gz |./tokenizer.perl –l en >corpusforRP/corpus_new4.tok.en .. (5.9)

The screenshot for execution of script 5.9 is given in Figure 5.5.

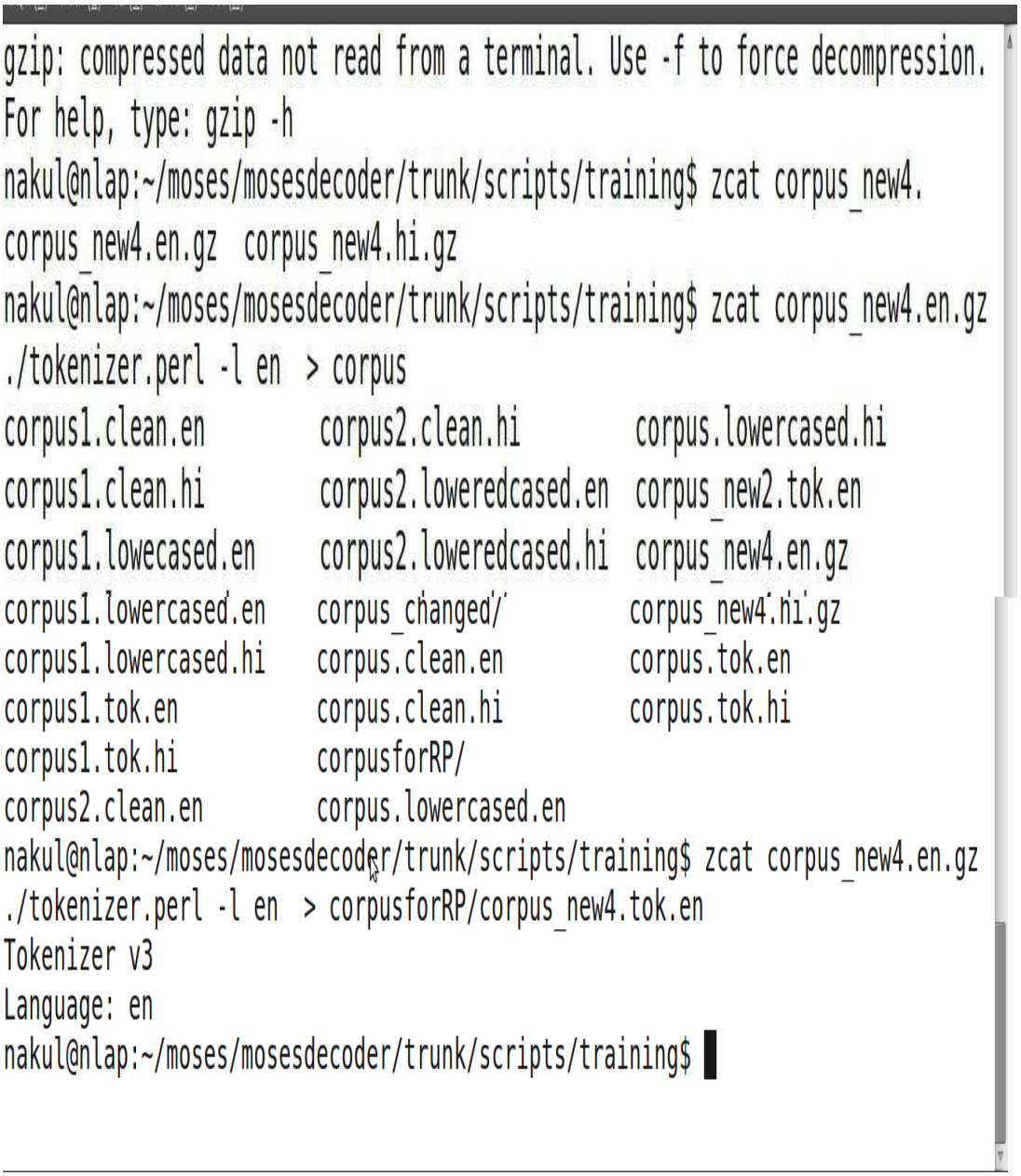


Figure 5.5: Tokenizing corpus

As a result of successful script execution, *corpus_new4.tok.en* is created with tokenized content.

5.8.2 Filtering out long sentences

Filtering out long sentences makes use of PERL script, *clean-corpus-n.perl*. The output of *tokenizer.perl* is accepted as input for *clean-corpus-n-perl.* This script removes long sentences from the corpus. It also removes redundant space characters and empty lines. Long sentences, are those which exceed word limit of 40 words. The system does not accept empty lines, hence they are removed. *GIZA++* takes very long time to train on long sentences. *Clean-corpus-n.perl* is used to reduce the length of sentences. The script is executed as given in 5.10.

./clean-corpus-n.perl corpusforRP/corpus_new4.tok en hicorpusforRP/corpus_ new4. clean 1 40 **--- (5.10)**

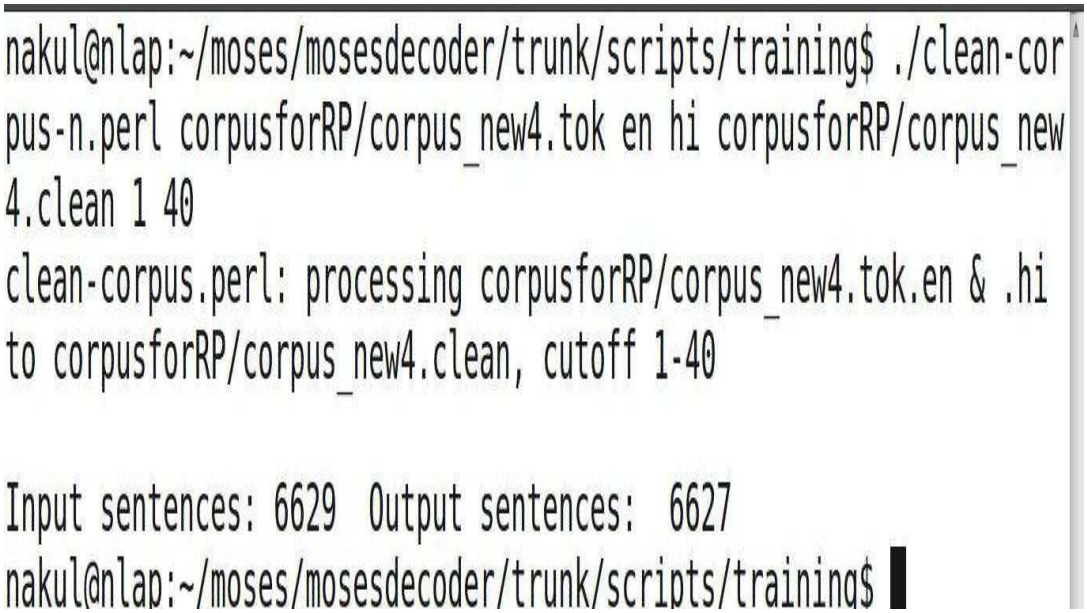


Figure 5.6: Filtering out long sentences

5.8.3 Lowercasing data

The data which is fed in for training the Moses software must be in small case. This is accomplished using *lowercase.perl*. Figure 5.6 shows a lowercased data. The script for lowercasing data is given in 5.11

./lowercase.perl <corpusforRP/corpus_new4.clean.en|more ... (5.11)

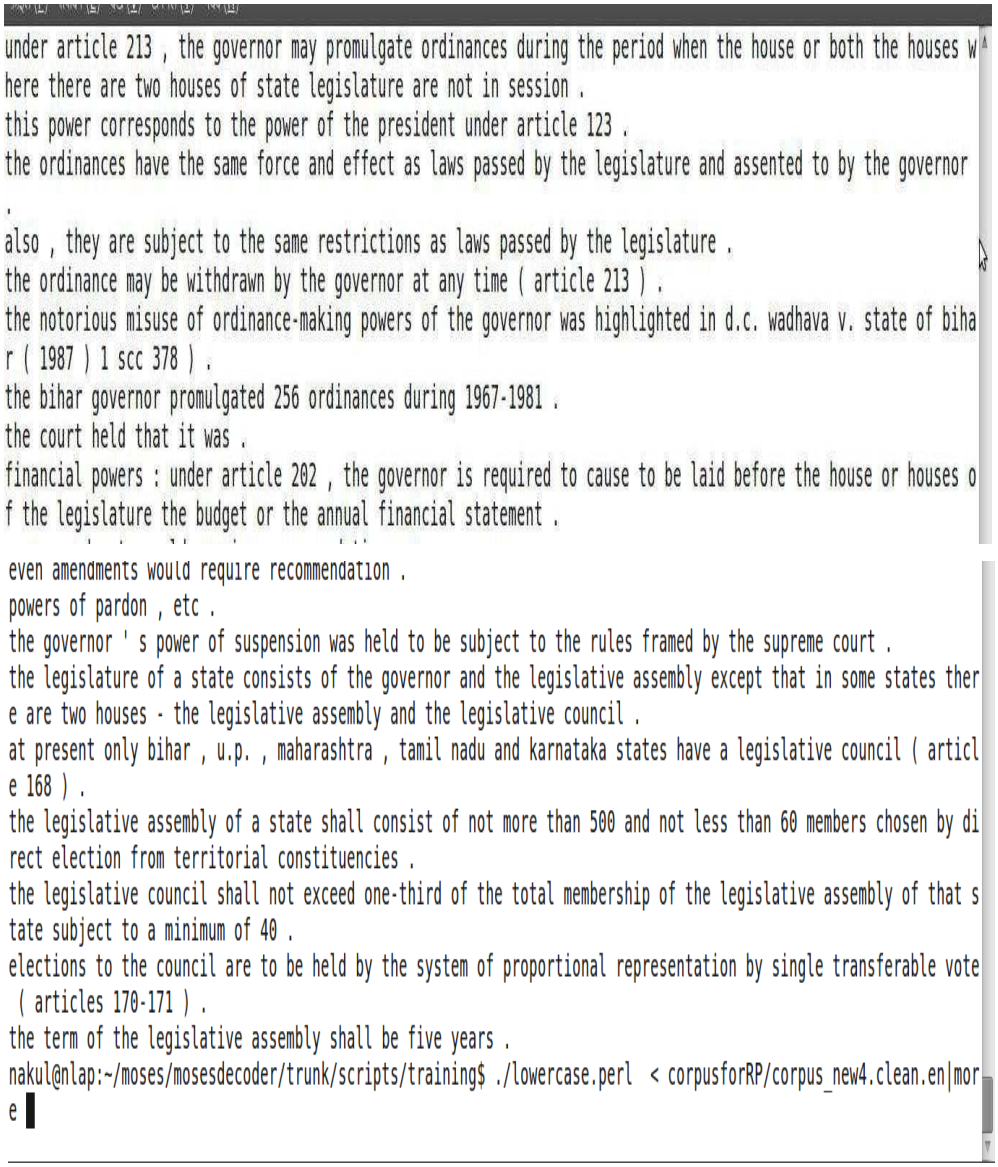

under article 213 , the governor may promulgate ordinances during the period when the house or both the houses w
here there are two houses of state legislature are not in session .
this power corresponds to the power of the president under article 123 .
the ordinances have the same force and effect as laws passed by the legislature and assented to by the governor
.
also , they are subject to the same restrictions as laws passed by the legislature .
the ordinance may be withdrawn by the governor at any time (article 213) .
the notorious misuse of ordinance-making powers of the governor was highlighted in d.c. wadhava v. state of biha
r (1987) 1 scc 378) .
the bihar governor promulgated 256 ordinances during 1967-1981 .
the court held that it was .
financial powers : under article 202 , the governor is required to cause to be laid before the house or houses o
f the legislature the budget or the annual financial statement .
even amendments would require recommendation .
powers of pardon , etc .
the governor ' s power of suspension was held to be subject to the rules framed by the supreme court .
the legislature of a state consists of the governor and the legislative assembly except that in some states ther
e are two houses - the legislative assembly and the legislative council .
at present only bihar , u.p. , maharashtra , tamil nadu and karnataka states have a legislative council (articl
e 168) .
the legislative assembly of a state shall consist of not more than 500 and not less than 60 members chosen by di
rect election from territorial constituencies .
the legislative council shall not exceed one-third of the total membership of the legislative assembly of that s
tate subject to a minimum of 40 .
elections to the council are to be held by the system of proportional representation by single transferable vote
(articles 170-171) .
the term of the legislative assembly shall be five years .
nakul@nlap:~/moses/mosesdecoder/trunk/scripts/training$ ./lowercase.perl < corpusforRP/corpus_new4.clean.en|mor
e

Figure 5.7: Lowercasing output

5.9 Generating Language Model

For Language Model (LM), SR International's LM model (SRILM) is used. SRILM is compatible with Moses decoder and GIZA++ Translation Model (TM). Language Model (LM)'s directory structure along with its description is shown in Table 5.1:

Table 5.1: Directory Structure of LM Model

Directory Structure	Descriptions
bin	Released Programs
lib	Released libraries
include	Released Header files
misc	Miscellaneous C and C++ convenience lib
Destruct	C++ data Structures
lm	Language Model libraries and tools

Ngram-count

Ngram-count counts the number of *n-gram* of the corpus. *Ngram-count* also builds the language model from the generated counts. The format of LM is also shown by ngram-format file.

The command for generating language model is given in 5.12

./ngram-count –order 3 –text corpus_new4.lowercased.hi –lm tamil.lm –write count.cnt … (5.12)

The description of parameters for PERL script, *ngram-count* is given in Table 5.2.

Table 5.2: Parameters of ngram-count

Parameter	**Description**
order	This parameter sets the maximal order of N-grams to count and the order of estimated LM. Default value is 3.

text	Generate *n-gram* counts from text file. Text file should contain one sentence unit per line. Begin/end sentence tokens are added if not already present. Empty lines are ignored.
write	Write count into mentioned file

The initial contents of the *tamil.lm* file created by *ngram-count* are shown in Figure 5.8.

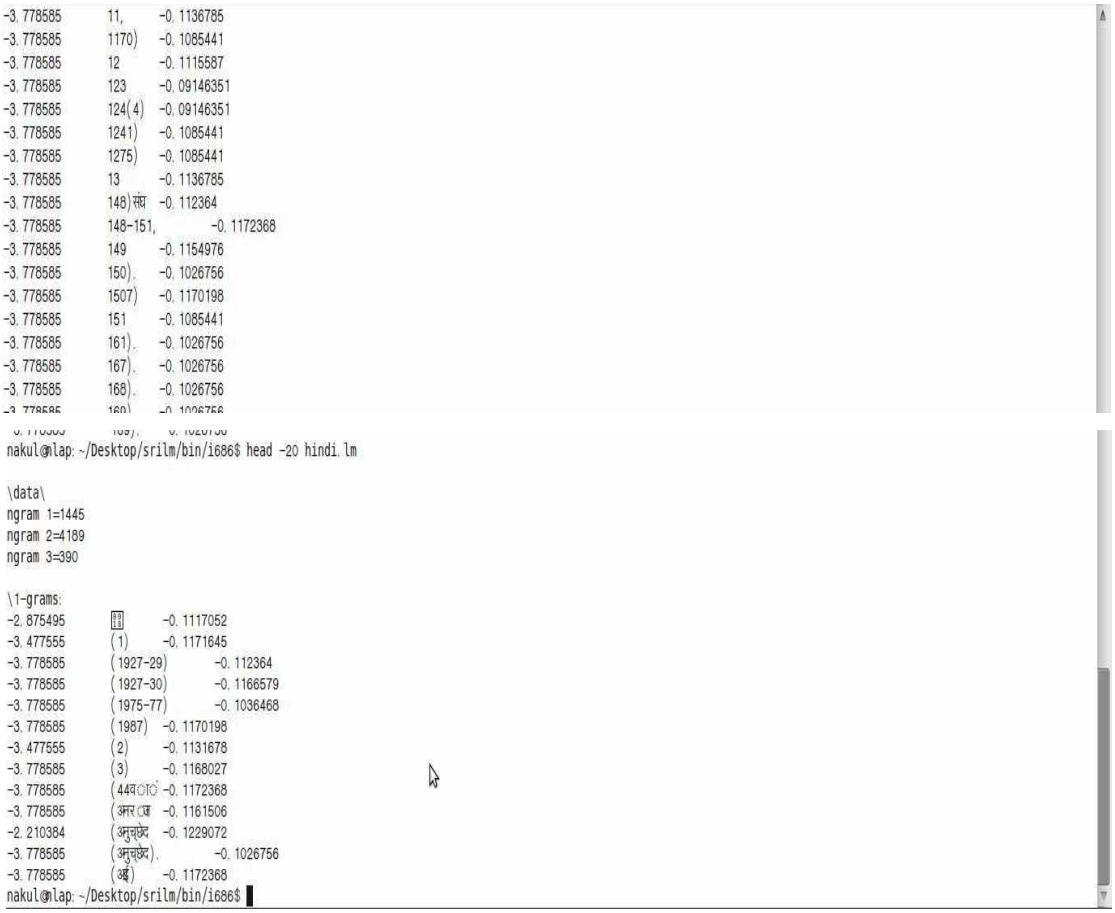


Figure 5.8: Contents of tamil.*lm* (in *ngram file format)*

The keyword *data*\ indicates the beginning of *lm* file. The total count of individual *ngrams*, found in the corpus is then mentioned after *data*\ keyword. For each n-gram (1-gram, 2-gram, *etc.*), there are individual sub-sections. Each sub-section starts with conditional probability of the *n-gram*. This probability is to the base of log 10. This is followed by the word which constitutes *n-gram*.

5.9.1 Installation of SRILM

The installation of SRILM involves following steps:

i). Unpack. It should give a top-level directory with the subdirectories listedin README, as well as a few documentation files and a Makefile.

ii). SRILM variable should then be set to the top-level Makefile. This path should be absolute starting from the root directory.

Specific to the architecture, the contents common/Makefile.machine.<platform> define the platform-dependent variables. The '*make*' command uses the dependencies in the Makefile to decide what parts of the program need to be compiled. The parameters are as shown in 5.13.

make MACHINE_TYPE=foo ... (5.13)

The variables in Makefile need to be changed are shown in Table 5.3.

Variable	**Changed value**
CC,CXX	This variable should be set to the complier or compiler version.
PIC_FLAG	This variable should be set to indicate the position-independent code.
DEMANGLE_FILTER	If program "c++filt" is not installed, this variable is set to empty.
TCL_INCLUDE, TCL_LIBRARY	These variables point to the location of Tool Command Language's (TCL) header files.

Following free third-party software's are also required to build SRILM:

- gcc version 3.4.3 or higher
- GNU make
- C shell (installed in /bin/csh)
- John Ousterhout's Tcl toolkit.

In the top-level directory, command 4.7, 4.8 are run to build SRILM.

gnumake World …. (5.14)

make World …. (5.15)

This will create the directories:

bin/
lib/
include/

bin directory stores the executable files of SRILM software. The released library files are stored in lib directory. The released header files are present in include directory.

5.10 Generating Translation Model

The software that aids in developing Translation Model is *GIZA++*. *GIZA++* is extension of *GIZA* software (http://www.clsp.jhu.edu/ws99/projects/mt/toolkit) which was developed at the Center for Language and Speech Processing at Johns-Hopkins University (CLSP/JHU). *GIZA++* includes a lot of additional features. The extensions of *GIZA++* were designed and written by Franz Josef Och.

GIZA++ has following features:

- Implements full IBM-4 alignment model
- Implements IBM-5: dependency on word classes, smoothing,
- Implements HMM alignment model
- Smoothing for fertility, distortion/alignment parameters
- Improved perplexity calculation for models IBM-1, IBM-2 and HMM.

The latest version of *Moses* software embeds calls to *GIZA++* and *mkcls* software's, hence there no need to call them separately.

5.10.1 Installation of *GIZA++*

In order to compile *GIZA++*, g++ compiler version 3.3 or higher is needed. Some changes are required to be made in the Makefile of GIZA directory as follows: The variables in Makefile of *GIZA++* directory need to changed, shown in Table 5.4.

Table 5.4: Variables in Makefile of *GIZA++* to be changed

Variable	Changed Value
CXX	This variable should indicate to version of g++ complier.
Opt	GIZA++ snt2plain.out plain2snt.out snt2cooc.out

GIZA++ is installed by issuing command given in 5.16.

$ make GIZA++ … (5.16)

5.11 Generating Decoder

Moses software helps in decoding stage of SMT. It allows us to train translation models for any language pair. The pre-requisite for the translation is already translated, parallel corpus.

5.11.1 Installation of Moses

Moses can be got from any svn repository. Before installing Moses, which is the statistical decoder for SMT, corresponding LM and TM tools must be installed. For LM, installation and compilation of SRILM must be done and for TM installation and compilation of GIZA++ must be done. Following compatible libraries are needed on UNIX system for running the SRILM software.

- A template-capable ANSI-C/C++ compiler, gcc version 3.4.3 or higher
- GNU make, to control compilation and installation.

- GNU gawk, required for many of the utility scripts.
- GNU gzip to unpack the distribution and to allow SRILM programs to handle compressed data files.
- The Tcl embeddable scripting language library.

These are installed by issuing the command as given in 5.17

.

$> sudo apt-get install g++ make gawk gzip tcl8.4 tcl8.4-dev … (5.17)

The Makefile in the SRILM is changed as shown in Table 5.5.

Table 5.5: Variables to be changed in Makefile

Variable	Changed value
SRILM	This variable must point to the SRILM's home directory.
MACHINE_TYPE	This variable points to the architecture of the system (i686, i386).
CC	/usr/bin/gcc$(GCC_FLAGS)
CXX	/usr/bin/g++$(GCC_FLAGS)-DINSTANTIATE_TEMPLATES
TCL_LIBRARY	/usr/lib/libtcl8.4.so
TCL_INCLUDE	/usr/include/tcl8.4/

After changing the Makefile, compilation of *Moses* is done command given in 5.18:

$ sudo make …(5.18)

If no error come, then the command in 5.19.is run.

$sudo make World … (5.19)

Some of the extra packages which need to be installed are done by issuing command
mentioned in 5.20.

$ sudo apt-get install autoconf automake texinfo zlib1g zlib1g-dev
zlib-bin zlibc ... (5.20)

The makefiles are regenerate as given in 5.21 to 5.24.

$ cd ~/mosesdecoder… … (5.21)
$ ./regenerate-makefiles… … (5.21)

Configuration for compilation is done as:

$ ln -s $SRILM …. … (5.21)
$ env LDFLAGS=-static && ./configure --with-srilm=$SRILM… … (5.24)

and compile:

$ make –j 4 ... (5.25)

5.11.2 Training Moses decoder

Moses toolkit embeds calls to Translation Model (GIZA++) software inside its training script. As a result, the phrase and reordering table get created. The script that does this is called train-factored-model.perl. Training of Moses decoder is done in nine steps. These are as follows.

- Prepare data
- Run GIZA++
- Align words
- Get lexical translation table
- Extract phrases
- Score phrases

□□Build lexicalized reordering model

□□Build generation models.

□□Create configuration file

The preparation of data (corpus) for this is already discussed in the earlier sections. The executable of train-factored-model is called as given in 5.26. Table 5.6 gives explanation of the parameters of training Moses.

./train-factored-phrase-model.perl -scripts-root-dir /home/nakul/moses/mosesdecoder/trunk/scripts/training/moses-scripts/scripts-20110405-1055/ -root-dir . --corpus corpus_new5.loweredcased -f en -e hi -lm 0:3:/home/nakul/moses/mosesdecoder/trunk/scripts/training/moses-scripts/scripts-20110405-1055/training/Tamil_lm5.lm>& training_new5.out & …(5.26)

Table 5.6: Parameters for training Moses

Arguments	**Description**
scripts-root-dir	The directory of Moses scripts which was created by doing make release.
corpus	Specifies the corpus files which are fed as input for undertaking training.
f	Source language corpus, from which translation will be done.
e	Target language corpus, into which translation will be done.
lm	Path to the Language Model file.

5.11.3 Tuning Moses decoder

The Moses software makes use of weights given in moses.ini to translate text. The default weights are generated by the system during its training. These weights are present in moses.ini, which is the configuration file of Moses. The most important

part is tuning of model parameters set in Moses.ini file. The quality of translation is improved, which is done by using PERL script (mert-moses.perl). The syntax of this command is given in 5.27.

.

./mert-moses.pl corpus_new5.loweredcased.en corpus_new5.loweredcased.hi model/moses.ini --working-dir /home/nakul/moses/mosesdecoder/trunk/mert/ --rootdir /home/nakul/moses/mosesdecoder/trunk/scripts/training/moses-scripts/scripts-20110405-1055/ --decoder-flags "-v 0" >& mert2.out& …(5.27)

Table 5.7: Parameters of mert-moses.pl

Arguments	**Description**
working-dir	The directory where all files will be created. This is the path to mert's directory
root-dir	This switch refers to the main directory inwhich system is working.
decoder-flags	This is a extra parameters for the decoder

The contents of mert2.out get updated as the script gets executed. Table 5.7 gives the explanation of parameters in tuning Moses.

5.11.4 Running Moses decoder

The Moses decoder's executable file is present in directory '*/home/nakul/mosesdecoder/trunk/moses-cmd/src/moses*'. The essential parameter required to run Moses, is the path to configuration file of Moses (Moses.ini).

The script 5.28 allows Moses decoder to run in interactive mode. The English language sentence is given as input and corresponding result in Tamil is produced.

./moses -f ~/moses/mosesdecoder/trunk/scripts/training/moses-scripts/scripts-20110405-1055/training/model/moses.ini . .... (5.28)

Figure 5.9 shows Moses decoder running in an interactive mode.

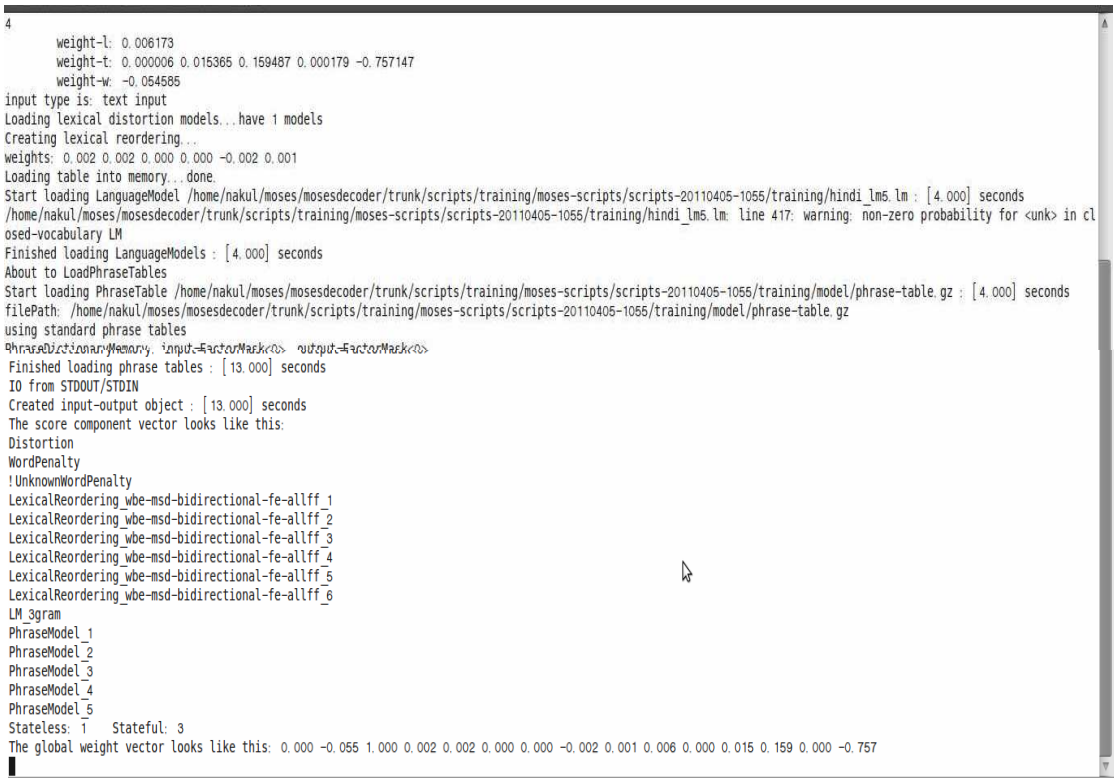


Figure 5.9: Interactive mode of Moses

Consider an English sentence ‘how are you?’ Moses decoder accepted this input in the interactive mode. The result of this translation is shown in Figure 5.8.

Figure 5.10 Result of English sentence ‘how are you?’

By executing Moses in interactive mode, 90 sentences were translated to Tamil language. Table 5.8 gives the English sentences along with the corresponding translation done by Moses into Tamil language.

Table 5.8: English to Tamil sentences generated by Moses

Sno	Input English Sentence	Output Tamil Sentence generated by the system
1	I play foot ball daily evening.	நான் தினமும் மாலையில் கால்பந்து விளையாடுகிறேன்.
2	we bought a new scooter last month.	நாங்கள் சென்ற மாதம் ஒரு புதிய ஸ்கூட்டர் வாங்கின்னோம்.

3	Two birds are flying in the sky.	வானத்தில் இரண்டு பறவைகள் பறந்துகொண்டிருக்கின்றன.
4	Doctor will come to this ward morning 6 o' clock	டாக்டர் இந்த வார்டுக்குக் காலை 6 மணிக்கு வருவார்.
5	The Chief Minister will speak in the crowd.	முதலமைச்சர் அந்தக் கூட்டத்தில் பேசுவார்.
6	My sister might win the first prize in the music competition.	எனது தங்கை இசைப் போட்டியில் முதற் பரிசை வெல்லக் கூடும்.
7	Someone has stolen my wrist watch.	யாரோ ஒருவர் என்னுடைய கைக்கடிகாரத்தைத் திருடிவிட்டார்.
8	The Cholan Express arrives to Thanjavur evening 6 o' clock	சோழன் எக்ஸ்பிரஸ் மாலை 6 மணிக்குத் தஞ்சாவூருக்கு வந்துசேர்கிறது.
9	You should finish this work within this evenin 5.30.	நீங்கள் இந்த வேலையை இன்று மாலை 5.30 மணிக்குள் முடிக்கவேண்டும்.
10	Mr.Kalai is going to become a minister.	திரு.கலை ஓர் அமைச்சர் ஆகப்போகிறார்.
11	Let the Child sleep on its mother's lap.	அந்தக் குழந்தை அதன் தாயின் மடியின் மேல் தூங்கட்டும்.
12	I had already read the book.	நான் ஏற்கனவே அந்தப் புத்தகத்தைப் படித்திருந்தேன்.
13	That girls might be going to the temple.	அந்த பெண்கள் கோவிலுக்குச் சென்றுகொண்டிருக்கக்கூடும்.
14	That child has fallen in to the river.	அந்தக் குழந்தை ஆற்றுக்குள் விழுந்துவிட்டது.
15	You should do excerises daily.	நீங்கள் தினமும் உடற்பயிற்சிகள் செய்யவேண்டும்.
16	He has been suffering from blood pressure for the past two years.	அவர் கடந்த 2 ஆண்டுகளாக இரத்த அழுத்தத்தால் கஷ்டப்பட்டுவருகிறார்.
17	You are wasting your time.	நீங்கள் உங்கள் காலத்தை வீணாக்கிக்கொண்டிருக்கிறீர்கள்.
18	That boys might be playing basketball in the playground.	அந்தப் பையன்கள் விளையாட்டரங்கில் கூடைப்பந்து விளையாடிக்கொண்டிருக்கக்கூடும்.

19	I can repair this wrist watch.	என்னால் இந்தக் கைக்கடிகாரத்தைப் பழுதுபார்க்கமுடியும்.
20	Ibramhim might reach Madurai by this time.	இப்ராஹிம் இந்நேரம் மதுரையை அடைந்த்ருக்கக்கூடும்.
21	I had to take my grandmother to the hospital.	நான் என் பாட்டியை மருத்துவமனைக்கு அழைத்துச்செல்லவேண்டியிருந்தது.
22	The postman will be coming this way morning 7 o'clock.	தபால்காரர் காலை 7.00 மணிக்கு இவ்வழியாக வந்துகொண்டிருப்பார்.
23	Our party might come to power in Tamilnadu.	எங்களுடைய கட்சி தமிழ்நாட்டில் ஆட்சிக்கு வரக்கூடும்.
24	My mother is teaching english to me.	என்னுடைய தாயார் எனக்கு ஆங்கிலம் கற்பித்துக்கொண்டிருக்கிறார்.
25	Mugan can speak english fluently.	முருகனால் நன்றாக ஆங்கிலம் பேசமுடிகிறது.
26	Thiruvalluvar is the author of Thirukkural.	திருக்குறளின் ஆசிரியர் திருவள்ளுவர்.
27	The price of the coconut oil is Rs.40 per liter.	தேங்காய் எண்ணெயின் விலை லிட்டருக்கு ரூ.40.
28	He is a honest man.	அவர் ஒரு நேர்மையான மனிதர்.
29	Mr.Mani is the collector of Erode.	திரு.மணி ஈரோடு மாவட்டதின் கலெக்டராக இருக்கிறார்.
30	This house is very comfortable.	இந்த வீடு மிகவும் வசதியாக இருக்கிறது.
31	The elephant's color is black.	யானையின் நிறம் கறுப்பு.
32	My father was ill yesterday.	நேற்று என்னுடைய தகப்பனார் உடல் நலமில்லாமல் இருந்தார்.
33	It is rice. It was paddy. It will be food.	இது அரிசியாக இருக்கிறது. இது நெல்லாக இருந்தது. இது சோறாக இருக்கும்.
34	I will become a rich man within five years.	ஐந்து ஆண்டுகளில் நான் ஒரு பணக்காரனாக ஆவேன்.
35	You should be very careful.	நீங்கள் மிகவும் கவனத்தோடு இருக்கவேண்டும்.
36	It is a difficult question.	இது ஒரு கடினமான கேள்வி.

37	I will become a manager of this company in 2005.	2005-இல் நான் இந்தக் கம்பெனியின் மேலாளராக இருப்பேன்.
38	The peacock is the national bird of India.	இந்தியாவின் தேசியப் பறவை மயில்.
39	The Cricket is the national game of England.	கிரிக்கெட் இங்கிலாந்தின் தேசிய விளையாட்டு.
40	Jeniva is the capital of Switcherland.	சுவிட்சர்லாந்தின் தலைநகரம் ஜெனீவா.
41	Your mother was very kind.	உன்னுடைய தாயார் மிகவும் அன்பானவராக இருந்தார்.
42	The price of the parker pen is Rs. 160.	பார்க்கர் பேனாவின் விலை ரூ.160/-
43	He is a retired judge of hight court.	அவர் ஓர் ஓய்வுபெற்ற உயர்நீதி மன்ற நீதிபதி.
44	He is a luck man.	அவர் ஓர் அதிர்ஷ்டசாலி.
45	Your futuer will be bright.	உன்னுடைய எதிர்காலம் மிகவும் ஒளிமயமானதாக இருக்கும்.
46	Your mother calls you.	உன்னுடைய தாயார் உன்னை அழைக்கிறார்.
47	I will go to chennai tomorrow.	நான் நாளை சென்னைக்குச் செல்வேன்.
48	The evening show starts 6 P.M.	மாலைக் காட்சி 6.00 மணிக்குத் தொடங்குகிறது.
49	I will buy a new scooter next month.	நான் அடுத்த மாதம் ஒரு புதிய ஸ்கூட்டர் வாங்குவேன்.
50	The Temple bell rings morning 6 o' clock.	கோவில் மணி காலை 6.00 மணிக்கு ஒலிக்கிறது.
51	Police caught thieves.	போலீசார் திருடர்களைப் பிடித்தார்கள்.
52	I ate two idlies in this morning.	நான் இன்று காலையில் இரண்டு இட்லிகள் மட்டும் சாப்பிட்டேன்.
53	I play foot ball daily morning.	நான் தினமும் காலையில் கூடைப் பந்து விளையாடுகிறேன்.
54	My father earns Rs.15000 per month.	என்னுடைய தகப்பனார் மாதமொன்றுக்கு ரூ.15000/- சம்பாதிக்கிறார்.

55	The milkman comes morning 5 o' clock. My mother prepares coffee at 5.30 A.M.	பால்காரர் காலை 5.00 மணிக்கு வருகிறார். என்னுடைய தாயார் காலை 5.30 மணிக்குத் காபி தயாரிக்கிறார்.
56	The principal and the student's leader will receive the chief guest at airport.	முதல்வரும் மாணவர் தலைவரும் தலைமை விருந்தினரை விமான நிலையத்தில் வரவேற்பார்கள்.
57	The magic man will fly in the sky.	அந்த மந்திரவாதி வானத்தில் பறப்பான்.
58	I always use mysoore sandal soap.	நான் எப்போதும் மைசூர் சந்தன சோப்பைப் பயன்படுத்துகிறேன்.
59	This train reaches new Delhi at 11.45 P.M.	இந்த ரயில் இரவு 11.45 மணிக்குப் புதுடில்லியை அடைகிறது.
60	I put the book on the table.	நான் அந்தப் புத்தகத்தை அந்த மேசையின் மேல் வைத்தேன்.
61	We will spend two weeks in Ooty.	நாங்கள் ஊட்டியில் 2 வாரங்களைக் கழிப்போம்.
62	We will stay in Ooty two weeks.	நாங்கள் ஊட்டியில் 2 வாரங்கள் தங்குவோம்.
63	Doctor gives medicines to patients.	டாக்டர் நோயாளிகளுக்கு மருந்துகள் கொடுக்கிறார்.
64	That old lady sells flowers. She earns Rs.50 per day.	அந்த மூதாட்டி பூக்கள் விற்கிறாள். அவள் நாளொன்றுக்கு ரூ.50/- சம்பாதிக்கிறாள்.
65	A bird flies in the sky. Birds fly in the sky.	வானத்தில் ஒரு பறவை பறக்கிறது.வானத்தில் பறவைகள் பறக்கின்றன.
66	Gopal wrote a letter to me in English. I answerd him in English.	கோபால் எனக்கு ஆங்கிலத்தில் ஒரு கடிதம் எழுதினான். நான் அவனுக்கு ஆங்கிலத்தில் பதிலளித்தேன்.
67	My father got a loan from the Indian Bank. He built a house in Arul Nagar.	என்னுடைய தகப்பனார் இந்தியன் வங்கியிலிருந்து ஒரு கடன் பெற்றார். அவர் அருள் நகரில் ஒரு வீடு கட்டினார்.
68	I will buy a safari suit for my birthday.	நான் என்னுடைய பிறந்த நாளுக்காக ஒரு ச஻பாரி சூட் வாங்குவேன்.
69	I will send new year greetings to my friends.	நான் என் நண்பர்களுக்குப் புத்தாண்டு வாழ்த்துகள் அனுப்புவேன்.

70	My sister won the first prize in the music competition.	என்னுடைய தங்கை இசைப் போட்டியில் முதற் பரிசை வென்றாள்.
71	India took 120 runs before the lunch break.	இந்தியா பகலுணவு இடைவேளைக்கு முன்னர் 120 ரன்கள் எடுத்தது.
72	India got freedom in 1947.	இந்தியா 1947-இல் விடுதலை பெற்றது.
73	My friend muthu will marry tamil cini actress poongodi next month	என் நண்பன் முத்து அடுத்த மாதம் தமிழ்த் திரைப்பட நடிகை பூங்கொடியை மணந்துகொள்வான்.
74	The Corporation supplies drinking water to this by lorry.	நகராட்சி இந்தத் தெருவுக்கு லாரி மூலம் குடிதண்ணீர் வழங்குகிறது.
75	We cultivate sugarcane in our fields.	நாங்கள் எங்கள் வயல்களில் கரும்பு பயிரிடுகிறோம்.
76	The price of the petrol will increase soon.	பெட்ரோலின் விலை விரைவில் உயரும்.
77	The sun set in the west.	சூரியன் மேற்கில் மறைகிறது.
78	Students threw stones at the bus.	மாணவர்கள் அந்தப் பேருந்தின் மீது கற்களை வீசினார்கள்.
79	We see stars at night in the sky.	நாம் இரவில் வானத்தில் நட்சத்திரங்களைப் பார்க்கிறோம்.
80	I deposited Rs.10000 in a bank before five years. I will get back Rs.20000 next year.	நான் ஒரு வங்கியில் 5 ஆண்டுகளுக்கு முன்னர் ரூ.10,000/- டெபாசிட் பண்ணினேன். நான் அடுத்த ஆண்டு ரூ.20,000/- திரும்பப்பெறுவேன்.
81	I resigned my job.	நான் என்னுடைய பதவியை ராஜினாமாசெய்தேன்.
82	That mad man murdered three members with a small knife.	அந்தப் பைத்தியக்காரன் ஒரு சிறிய கத்தியைக் கொண்டு 3 பேர்களைக் கொலைசெய்தான்.
83	He donates blood on his brithday everyday.	அவன் ஒவ்வோராண்டும் தன்னுடைய பிறந்தநாளன்று இரத்ததானம்செய்கிறான்.
84	The people of Tamilnadu celebrate pongal festival in the month of thai every year in very grand manner.	தமிழ்நாட்டு மக்கள் ஒவ்வோராண்டும் தை மாதத்தில் பொங்கல் விழாவை மிகச் சிறப்பான முறையில் கொண்டாடுகிறார்கள்.

85	150 countries participate in the olympic games this time.	இந்தத் தடவை நூற்றைம்பது நாடுகள் ஒலிம்பிக் விளையாட்டில் பங்கெடுத்துகொள்கின்றன.
86	I attended in my friedn's marriage.	நான் என் நண்பனுடைய திருமணத்தில் கலந்துகொ ண்டேன்.
87	India defeated south africa in the final match.	இறுதி ஆட்டத்தில் இந்தியா தென்னாப்பிரிக்காவைத் தோற்கடித்தது.
88	The principal dismissed three students from the college. They misbehaved with students.	முதல்வர் மூன்று மாணவர்களைக் கல்லூரியிலிருந்து நீக்கினார். அவர்கள் மாணவியர்களிடம் தவறாகநடந்துகொண்டார்கள்.
89	Our college reopens fist of june month.	எங்கள் கல்லூரி ஜூன் மாதம் முதல் தேதி திறக்கிறது.
90	She helps to her mother in coooking.	அவள் தன் தாயாருக்குச் சமையலில் உதவுகிறாள்.
91	About 25 lakhs tourists visit to India every year.	சுமார் 25 லட்சம் சுற்றுலாப் பயணிகள் ஒவ்வோராண்டும் இந்தியாவுக்கு வருகைதருகிறார்கள்.
92	About 30000 birds arrive to birds sanctuary every year.	ஒவ்வோராண்டும் சுமார் 30,000 பறவைகள் வேடந்தாங்கல் பறவைகள்புகலிடத்துக்கு வந்துசேர்கின்றன.
93	Muslims fasting in the month of Ramjan.	முஸ்லீம்கள் இரம்சான் மாதத்தில் உண்ணாநோன்பு நோற்கிறார்கள்.
94	Factories, buses and cars pollute the air.	தொழிற்சாலைகளும் பேருந்துகளும் கார்களும் காற்றை மாசுபடுத்துகின்றன.

5.12 EXPERIMENTAL FRAMEWORK

5.12.1 English–Tamil Phrase Based Statistical Machine Translation System

Tamil, a Dravidian language, is spoken by around 72 million people and is the official language of Tamil Nadu state government of India. Many resources in English are manually translated to Tamil, which consumes more time, human resource and

cost. Here a machine translation system based on the statistical approach for English to Tamil translation has been designed and implemented, in order to translate faster and cheaper.

5.12.2 Proposed System Architecture

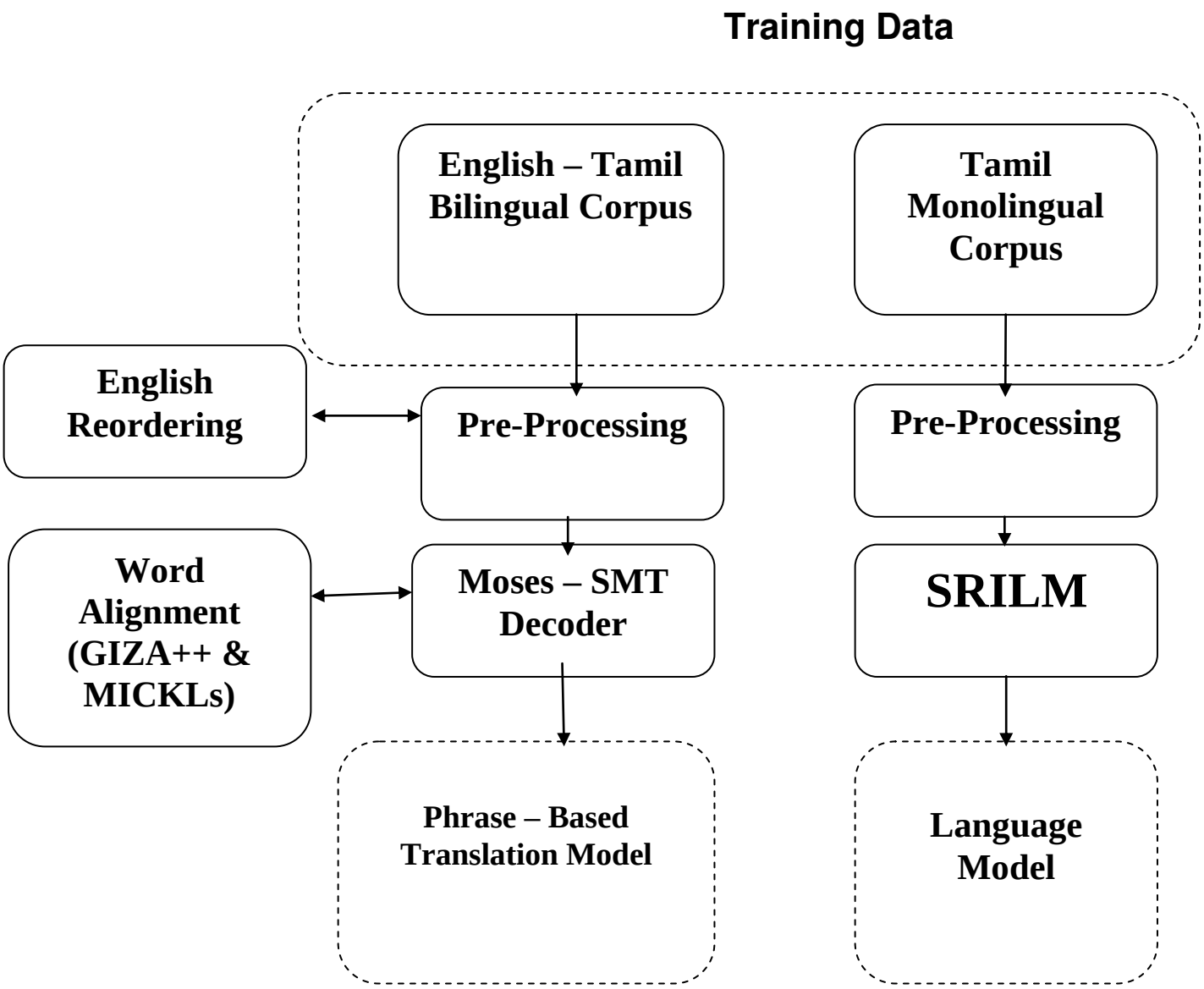


Fig. 5.12 System architecture of the training Phase of the phrase based English – Tamil statistical machine translation system

Fig. 4.2 System architecture of the testing phase of the phrase based English – Tamil statistical machine translation system

English is a highly positional language with rudimentary morphology, and default sentence structure as SVO. Indian languages are highly inflectional, with a rich morphology, relatively free word order, and default sentence structure as SOV. In addition, there are many stylistic differences. For example, it is common to see very long sentences in English, using abstract concepts as the subjects of sentences, and stringing several clauses together. Such constructions are not natural in Indian languages, and lead to major difficulties in producing good quality translations. Here, an external module to reorder the English sentence which is of the SVO word pattern to SOV word pattern, as SOV pattern is used in Tamil more often, has been integrated with pre-processing phase of the baseline system so has to train or test the English sentences after reordering. System Architecture of the training and testing phase of the phrase based English – Tamil statistical machine translation system is given in Figures 4.1 and 4.2 respectively.

5.13 Implementation

5.13.1 Development of Bilingual Corpus for English –Tamil language pair

The availability of parallel sentences for English-Tamil language pair is available, but not abundantly. In European countries, parallel data for many European language pair are available from the proceedings of the European Parliament. But in case of Tamil, no such parallel data are readily available. Hence English sentences have to be collected and manually translated to Tamil in order to create a bilingual corpus for English-Tamil language pair. Even though, if parallel data are available for English-Tamil language pair, there are chances that it might not be aligned properly and have to be separate the paragraphs in to individual sentences. This will employ a lot of human resource. This is a time extensive work

and has it is the main resource for the statistical machine translation system, more time and importance has to be provided in developing a bilingual corpus for English-Tamil language pair. During manual translations of English sentences to Tamil, terminology data banks for English-Tamil language pair are found to be very useful for humans.

5.13.2 Development of Monolingual Corpus for Tamil language

The situation for developing bilingual corpus for English-Tamil language pair is not the same for the development of monolingual corpus for Tamil language. Tamil data is available in the form of news in many websites of Tamil newspapers. And so it is not a tedious job to develop a monolingual corpus for Tamil language. But some human resource is necessary to perform some pre-processing to remove unnecessary words or characters from the data, manually.

5.13.3 Pre-processing of Corpus

Before providing the bilingual corpus of English-Tamil language pair and monolingual corpus of Tamil language to the statistical machine translation decoder Moses and the language modelling kit, SRILM, respectively for training the system in order to create translation models and language models, both the corpus has to be tokenized in order to separate the words and punctuations i.e.,'coming,' will be separated as ‗coming' and ‗,' with space in between them, lowercased in order to consider all the same words but differs in case has a single word (for example, 'He' and 'he' if not lowercased will be considered as different entities by the statistical systems which will be a problem whereas if lowercased this problem can be avoided) and in some cases clean the corpus so has remove the sentences from the corpus that exceeds the limit which is the maximum length of the parallel sentences to be considered in the corpus. Cleaning the corpus is not necessary in case of monolingual corpus of Tamil language.

5.13.4 Building Language Model

SRILM language modelling kit can be used to build an n-gram language model from the monolingual corpus of Tamil language. A script, ‗ngram-count', in SRILM can be used to generate n-gram language models of any order by specifying optional parameters such as interpolation, modified Kneser-Ney smoothing, absolute discounting, Good – Turing smoothing and Written-Bell smoothing for unseen n-grams. The output of this script will be a language model file that contains the n-gram probabilities of each word in the monolingual corpus. The general syntax of executing the script ‗ngram-count' in SRILM is,

> ngram-count -order n -[options] -text CORPUS_FILE –lm LM_FILE

Where,

order n - the order of the n-gram language model can be mentioned here, with ‗– order n', where ‗n' denotes the order of the n-gram model.

[options] – various switches, such as interpolate, kndiscount, ndiscount and so on, that can be used to generate the language model file.

text – the file name of the monolingual corpus file

lm – the file name of the language model file to be created by the script.

5.13.5 Building Phrase-Based Translation Model

To build a phrase-based translation model, the perl script, 'train-model.perl' in Moses is used. The train-model perl script involves the following steps,

- *Prepare the data:* convert the parallel corpus into a format that is suitable to GIZA++ toolkit. Two vocabulary files are generated and the parallel corpus is converted into a numbered format. The vocabulary files contain words, integer word identifiers and word count information. GIZA++ also requires words to be placed into word classes. This is done automatically calling the mkcls program. Word classes are only used for the IBM reordering model in GIZA++.

- *Run GIZA++:* GIZA++ is a freely available implementation of the IBM Models. We need it as an initial step to establish word alignments. Our word alignments are taken from the intersection of bidirectional runs of GIZA++ plus some additional alignment points from the union of the two runs. Running GIZA++ is the most time consuming step in the training process. It also requires a lot of memory. GIZA++ learns the translation tables of IBM Model 4, but we are only interested in the word alignment file.

- *Aligning words:* To establish word alignments based on the two GIZA++ alignments, a number of heuristics may be applied. The default heuristic growdiag- final starts with the intersection of the two alignments and then adds additional alignment points. Other possible alignment methods are intersection, grow, grow-diag, union, srctotgt and tgttosrc. Alternative alignment methods can be specified with the switch alignment.

- *Get lexical translation table:* Given the word alignment, it is quite straightforward to estimate a maximum likelihood lexical translation table. We estimate the $w(e \mid f)$ as well as the inverse $w(f \mid e)$ word translation table.

- *Extract Phrases:* In the phrase extraction step, all phrases are dumped into one big file. The content of this file is for each line: foreign phrase, English phrase, and alignment points. Alignment points are pairs (English, Tamil). Also, an inverted alignment file extract.inv is generated, and if the lexicalized reordering model is trained (default), a reordering file extract.o.

- *Score Phrases:* Subsequently, a translation table is created from the stored phrase translation pairs. The two steps are separated, because for larger translation models, the phrase translation table does not fit into memory. Fortunately, we never have to store the phrase translation table into memory; we can construct it on disk, itself. To estimate the phrase translation probability $\varphi(e \mid f)$we proceed as follows: First, the extract file is sorted. This ensures that all English phrase translations for a foreign phrase are next to each other in the file. Thus, we can process the file, one foreign phrase at a

time, collect counts and compute φ (e | f) for that foreign phrase f. To estimate φ (f | e) , the inverted file is sorted, and then φ (f | e) is estimated for an English phrase at a time. Next to phrase translation probability distributions φ (f | e) and φ (e | f) , additional phrase translation scoring functions can becomputed, e.g. lexical weighting, word penalty, phrase penalty, etc. Currently, lexical weighting is added for both directions and a fifth score is the phrase penalty. Currently, five different phrase translation scores are computed. They are, phrase translation probability φ (f | e) , lexical weighting *lex*(f | e) , phrase translation probability φ (e | f) , lexical weighting *lex*(e | f) and phrase penalty (always exp(1) =□2.718).

- *Build Reordering model:* By default, only a distance-based reordering model is included in final configuration. This model gives a cost linear to the reordering distance. For instance, skipping over two words costs twice as much as skipping over one word. Possible configurations are msdbidirectional- fe (default), msd-bidirectional-f, msd-fe, msd-f, monotonicitybidirectional- fe, monotonicity-bidirectional-f, monotonicity-fe and monotonicity-f.

- *Build Generation model:* The generation model is built from the target side of the parallel corpus. By default, forward and backward probabilities are computed. If you use the switch generation-type single only the probabilities in the direction of the step are computed.

- *Creating Configuration file:* As a final step, a configuration file for the decoder is generated with all the correct paths for the generated model and a number of default parameter settings. This file is called model/moses.ini Thus, the phrase-based translation model has been generated.

5.13.6 Tuning

Prepare a gold standard bilingual corpus of English-Tamil language pair in order to tune the translation model generated by the decoder from the bilingual corpus of the language pair given for training. The ‗mert-moses.perl' script in moses

is used to tune the translation model and it will generate a tuned moses configuration file. The tuned moses configuration file will be used hereafter to translate English sentences to Tamil.

Table 4.1 Experimental results for phrase-based statistical machine translation system

System	BLEU	NIST
Baseline	24.32	5.72
Baseline + Reordering	31.62	6.41

5.13.7 Experimental Results

After the tuning the phrase-based statistical machine translation system with the gold standard corpus, the system have been tested with a set of sentences. The output of the system has been evaluated with the reference translations using BLEU and NIST score. The experimental results have been tabulated in Table 4.1 and the sample output of the translations generated by this system is tabulated in Table 4.2.

Table 4.2 Sample output of translations generated by the phrase-based statistical machine translation system

English Sentence	Translated Tamil Sentence
I am playing football.	நான் football விளையாடுகிறேன்.
My elder brother is coming	என் அண்ணா வருகிறான்.

5.14 Handling Idioms and Phrasal Verbs in Machine Translation

5.14.1 Overview

In this section, work on incorporating a technique to handle phrasal verbs and idioms for English to Tamil machine translation was discussed. While translating from English to Tamil, both phrasal verbs and idioms in English have more chances, to get translated to Tamil in wrong sense. This is because of the idioms or phrasal verbs that convey individual meaning for each word in it instead of conveying a single meaning by considering it as a group of words while translating from English to Tamil. This in turn affects the accuracy of the translation. The proposed technique is used to handle the idioms and phrasal verbs during the translation process and it increases the accuracy of the translation. The BLEU and NIST scores calculated before and after handling the phrasal verbs and idioms during the translation process show a significant increase in the accuracy of the translation. This technique, proposed for English to Tamil machine translation system, can be incorporated with machine translation system for English to any language.

5.14.2 Idioms and Phrasal Verbs in Machine Translation

Machine translation is an important and most appropriate technology for localization in a linguistically diverged country like India. The reason for choosing automatic machine translation rather than human translation is that machine translation is better, faster and cheaper than human translation. Many resources such as news, weather reports, books, etc., in English are being manually translated to Indian languages. Of these, News and weather reports from all around the world are translated from English to Indian languages by human translators more often. Human translation is slow and also consumes more time and cost compared to machine translation. Hence, there is a good scope for machine translation to overcome the human translation, in near future. There are machine translation systems that are being developed in order to translate from English to Indian languages. But there are problems that make these systems not able to produce a good translation of text fromEnglish to Indian languages. Here we incorporate the technique with English-Tamil machine translation system.

One of the problems in English-Tamil machine translation system is to handle the idioms and phrasal verbs. A phrasal verb, which is a combination of a verb and a

preposition or adverb, creates a meaning different from its constituent verb. It should not be translated by considering its constituent verb alone. Similarly an idiom, which is usually a group of words, conveys a peculiar meaning and cannot be predicted from the meaning of the constituent words. It should be handled as a single unit during the translation process. But the existing machine translation system handles the translation of a phrasal verb by translating the constituent verb in it and idiom by translating each constituent word in it. This makes idioms and phrasal verbs to have a great impact in the accuracy of English-Tamil machine translation system.

Here a technique that can be used to handle idioms and phrasal verbs which can increase the accuracy of English - Tamil translation, when incorporated with any existing English - Tamil machine translation system is described. The technique consists of two phases, analyzing phase and grouping phase. In analyzing phase, the given English sentence is analyzed to find whether it contains any phrasal verbs or idioms. In grouping phase, if the given sentence is found to contain a phrasal verb or an idiom, then it will be grouped into a single unit and it will be categorized with a special tag in order to denote it as the phrasal verb or idiom. This tag will be considered instead of the part-of-speech tag during the translation process. This approach can be used in both rule based and factored statistical machine translation with some modifications.

5.14.3 Phrasal Verbs and Idioms – An Overview

As described earlier, a phrasal verb is a combination of a verb and a preposition or adverb that creates a meaning different from its original constituent verb. Phrasal verbs can be broadly classified into two categories, transitive and intransitive.

A transitive phrasal verb can either be followed by an object or it can contain an object between the verb and preposition or adverb and this can be further classified into separable and inseparable. Separable transitive phrasal verbs are those in which the object is placed between the verb and the preposition or adverb. Inseparable transitive phrasal verbs are those in which the object is placed after the

preposition or adverb. Also there exist some transitive phrasal verbs that can be considered in both cases, separable and inseparable.

Though some transitive phrasal verbs can be both separable and inseparable, the phrasal verb should take only the separable form when the object is a pronoun. An intransitive phrasal verb should neither be followed by an object nor should it contain an object between the verb and preposition or adverb. Examples for the types of phrasal verbs are illustrated in Table 4.3.

Table 4.3 Types of phrasal verbs with examples

Type		Phrasal Verb	Meaning	Example
Transitive	Separable	Cut ∗ off	Interrupt someone while they were speaking	She cut him off while he was talking
	Inseparable	Look into +	Investigate	The police are lokking into the murder
	Separable / inseparable	Pass ∗ out +	Distribute	We need to pass these sweets out. (Separable) We need to pass out these sweets (Inseparable)
Intransitive		Pass away	Die	He passed away

** - Object in between, + - Object after the verb and preposition or adverb*

An idiom is usually a group of words whose meaning will be peculiar and cannot be predicted from the meanings of the constituent words. Also, it can be considered as an expression that is not readily analysable from its grammatical construction or from the meaning of its component parts. In other words, an idiom is an expression, word, or phrase whose sense means something different from what the words literally imply. In most cases when an idiom is translated, either its meaning is changed or it is meaningless. There are estimated to be at least 25,000

idiomatic expressions in the English language. An idiom is generally a colloquial metaphor a term requiring some foundational knowledge, information, or experience, to use only within a culture, where conversational parties must possess common cultural references. Therefore, idioms are not considered part of the language, but part of the culture. In linguistics, idioms are usually presumed to be figures of speech contradicting the principle of compositionality which states that the meaning of a complex expression is determined by the meanings of its constituent expressions. In general, idioms are based on pair of words, number, nationality, colour, etc. and are illustrated with examples in Table 4.4.

Table 4.4 Types of idioms with examples

5.14.4 Challenges in Handling Idioms and Phrasal Verbs

The main problem in existing machine translation system due to phrasal verbs and idioms is that a phrasal verb is translated by considering the constituent verb in it, instead of considering it as a single unit. For example, the sentence

"The minister *passed away*"

will be translated as

"amaiccar *thUram thErcciyataiwthAr"* (அமைச்சர் தூரம் தேர்ச்சியடைந்தார்)

instead of

"amaiccar *iyaRkai eythinAr"*

(அமைச்சர் இயற்கை எய்தினார்.).

Here, the phrasal verb is translated in such a way that instead of conveying its meaning as a single unit i.e., '*to die*', conveys the meaning as '*to pass*' by considering the constituent verb in it.

Similarly, an idiom is translated by considering the constituent words in it, instead of considering it as a single unit during the translation process from English to Tamil. For example, consider the sentence

"This work is *a piece of cake*"

will be translated as

“iwtha vElai inirottiyin oru pakuthiyAkum”

(இந்த வேலை இனிரொட்டியின் ஒரு பகுதியாகும்),

instead of

“iwtha vElai *eLithAnathu*”

(இந்த வேலை எளிதானது).

Here the idiom is translated in such a way that the translation conveys the literal meaning of constituent words in the idiom (i.e., ‘a piece of cake‘), instead of conveying the meaning, ‘easy‘ by considering it as a single unit in the sentence. These examples above show how phrasal verbs and idioms affect the accuracy of the translation system. As idioms cannot be analysed from its grammatical construction, handling the idioms in translation process becomes a challenging task. Since idioms and phrases are used more frequently in English language, it becomes necessary to handle the idioms during the translation from English to Tamil.

In order to handle these phrasal verbs and idioms, a collection of most frequently used phrasal verbs and idioms have to be collected and manually translated to Tamil in such a way that it should convey the exact meaning or sense of the phrasal verb or idiom when considered as a single unit in the sentence. Lexical dictionary for these phrasal verbs and idioms is created with the collected phrasal verbs and idioms and its equivalent translation in Tamil. This dictionary can be referred by the machine translation system, if required, to replace the phrasal verbs or idioms in English with its Tamil equivalent. While creating the lexical dictionary for phrasal verbs, the dictionary is created with root form of the phrasal verbs, so that all the inflections of the phrasal verbs can be handled in a way similar to that of verbs. For example, instead of ’*passed away*‘ its root form ‘*pass away*‘ is added to the lexical dictionary.

Also in order to handle the separable transitive phrasal verbs, some rules have to be coded such that in case of phrasal verbs which can be both separable and inseparable and if it have pronoun as the object, it should be handled as separable. Some of the phrasal verbs convey one meaning when they are transitive, which is entirely different from the meaning when they take intransitive form. For example, the phrasal verb ‗*show up*‘ gives the meaning ‗make someone seem

inferior‘ in transitive case and ‗arrive without prior notice‘ in intransitive case. These cases are handled by taking the object in consideration, so that it distinguishes the transitive and the intransitive form of the phrasal verb during the translation process.

5.14.5 Implementation

The general block diagram of proposed technique to handle the phrasal verbs and idioms during English-Tamil machine translation system is given in Figure 4.3. The input to this technique can be a sentence in case of rule based machine translation and bilingual and monolingual corpus for training and input sentences in case of statistical machine translation. Before providing the input to the machine translation system for further process, the input is passed to the first phase of the proposed technique, Phrasal verbs and Idioms Analyser.

Here the input is thoroughly analysed for any phrasal verbs or idioms in it, by looking up in the list of phrasal verbs and idioms collected. If any phrasal verb or idiom is found to be in the sentence then it is passed to the second phase of the technique, the grouping phase.

In the grouping phase, the words in the phrasal verb or idiom that is found to be in the input in the analyser phase are grouped together into a single unit and a special tag is assigned to it, so that this phrasal verb or idiom will be considered as a single unit during the whole translation process.

In the grouping phase, while grouping the words in the phrasal verb which is of transitive separable type, the object in between the verb and the preposition or adverb is moved after the preposition or adverb in it. For example, the sentence,

“She *cut him off* while he was talking” will be grouped as

will be grouped as

“She *cut-off him* while he was talking”

and will be translated as

“avan pEcikkoNtirukkum pozuthu avaL avanai kURukkittaL”

(அவன் பேசிக்கொண்டிருக்கும் பொழுது அவள் குறுக்கிட்டாள்),

as the phrasal verbs are handled in the way similar to verbs. Lexical dictionary with 900 idioms and 241 phrasal verbs have been created for idioms and phrasal verbs, separately

Fig. 4.3 General block diagram for the proposed technique to handle phrasal verbs and idioms in machine translation system

The above block diagram for the proposed technique can be integrated to any English-Tamil rule based machine translation system or to any English-Tamil statistical machine translation, with some modifications in the general technique. The following section will give a clear idea of how this technique can be used in rule based and factored statistical machine translation.

5.14.5.1 Rule Based Machine Translation System

In rule based machine translation system, the given English sentence annotated with lemma, part of speech tag, morphological and dependency information is passed to the first-phase of the technique, Phrasal verbs and Idioms analyser phase, before passing the sentence to the actual translation process. In this phase, the analyser checks for any phrasal verbs or idioms present in the given sentence. If found, the sentence is passed to the grouping phase, where the words that form the phrasal verb or idiom found in the analyser phase are grouped together as a single unit in the sentence and it is assigned with a special tag ‘PHV‘ for phrasal verbs and ‘IDM‘ for idioms along with the annotated part of speech tag information.

In case of phrasal verbs which take both transitive and intransitive form, the form of the phrasal verb is differentiated by the object following it or in between the verb and adverb or preposition. An asterisk symbol is added to the end of root of the phrasal verb, if it is intransitive. So that while translating, the two forms of the phrasal verb can be differentiated easily. For example, intransitive form of the phrasal verb '*show up*' will be changed to '*show-up**' which means '*arrive without prior notice*'. All other annotated information of the words grouped to form a single unit is also grouped in the sequence of the words as in the phrasal verb or idiom. During the translation process, the unit assigned with the special tag 'PHV' will be handled as verb indeed, but during lexical replacement of English to Tamil, instead of retrieving from the lexical dictionary for verb, some modification has to be made in the existing system so that it retrieves from lexical dictionary for phrasal verbs and for the words with the tag 'IDM', the lexical replacement has to be made from the lexical dictionary for idioms. The block diagram for the modified technique for English-Tamil rule based machine translation system is shown in Figure 4.4.

Fig. 4.4 Modified block diagram for the proposed technique to handle phrasal verbs and idioms in rule based English-Tamil machine translation system

5.14.5.2 Factored Statistical Machine Translation System

In the existing factored statistical machine translation system before the training phase the bilingual and monolingual corpus is pre-processed by the proposed technique to group the phrasal verbs and idioms in to a single unit. Here, the term factored means the corpus along with information such as lemma, part-of-speech tag and morphological information for each word in every sentence in the corpus.

The statistical machine translation decoder translates the sentences from English to Tamil by considering the factored information as translation factors. Here, the technique has been modified so that in the proposed technique's analyser phase, the English sentences are analysed for phrasal verbs or idioms. If found, in the grouping phase the phrasal verbs or idioms in English as well as its equivalent in Tamil are also grouped into a single unit.

Also the Tamil monolingual corpus has been analysed for phrasal verbs or idioms, and grouped into a single unit, if found any. And the part-of-speech category for phrasal verbs and idioms are assigned as _PHV‘ and _IDM‘ respectively.

The technique is applied in a similar way to the monolingual corpusAfter the grouping phase of the technique, the bilingual and monolingual corpus is passed to the training phase of the decoder. During the testing phase, the factored sentence is pre-processed by this technique and then passed to the decoder for translation.

The output of the decoder is given to the morphological generator to generate the final translated sentence. Figure 4.5 shows the block diagram for the modified technique for English-Tamil factored statistical machine translation system.

Fig. 4.5 Modified block diagram for the proposed technique to handle phrasal verbs and idioms in factored English-Tamil statistical machine translation system

5.14.6 Experimental Results

The machine translation system for English-Tamil has been tested and evaluated for four cases, (1) the baseline machine translation system, (2) the baseline machine translation system with the proposed technique to handle phrasal verbs, (3) the baseline machine translation system with technique to handle idioms and (4) the baseline machine translation system with technique to handle both phrasal verbs and idioms, in both the rule based and factored statistical machine translation system. Table 4.5 Comparison of translation results of machine translation system with and without the proposed technique to handle phrasal verbs and idioms

Phrasal Verbs Or Idioms		English	Output of Baseline System	Output of Baseline System with proposed technique
Phra	Account	He should account for his	அவன் அவனுடைய	அவன் அவனுடைய தவறுகளுக்கு

	for	mistakes	தவறுகளுக்கு எண்ணவேண்டும்	விளக்கமளிக்கவேண்டும்
	Call off	The meeting was called off	கூட்டம் அழைக்கப்பட்டது	கூட்டம் ரத்தானது
	Pass out	He passed the sweets out	அவன் தேர்ச்சியடை இனிப்பான	அவன் இனிப்புகளை வினியோகித்தான்
Idioms	Jack of all trades.	Arun is a jack of all trades	அருண் அணைத்து வர்த்தங்களுக்கும் ஒரு ஜேக்	அருண் ஒரு சகலகலா வல்லவன்
	A piece of cake	This job is a piece of cake	இந்த வேலை இனிரொட்டியின் ஒரு வேலையாகும்	இந்த வேலை எளிதானது.
	Smell a rat	I smell a rat on seeing him	நான் அவனை கண்டவுடன் ஒரு எலியை நுகர்ந்தேன்	நான் அவனை கண்டவுடன் சந்தேகமடைந்தேன்

The rule based machine translation system has been evaluated with a test data set of 500 sentences. The factored statistical machine translation system has been trained with English – Tamil bilingual corpus with 20,000 parallel sentences and a Tamil monolingual corpus of 50,000 sentences and has been evaluated with another test data set of 500 sentences. Both the systems have been evaluated for

the four cases with BLEU and NIST score and the results shows that incorporating this technique to handle idioms and phrasal verbs has increased the accuracy of the existing English - Tamil machine translation systems.

Comparison of how the sentences containing phrasal verbs or idioms in English gets translated to Tamil with the existing machine translation system and the existing machine translation system with the proposed technique to handle the phrasal verbs and idioms are illustrated with examples in Table 4.5.

5.14.3 Automated Factored Information Generation for English and Tamil

Phrase-based models do not consider linguistic information other than words. This linguistic information other than words should be considered, as with this information the quality of the translation will improve. Thus the idea of making use of the syntactic information in statistical machine translation resulted in factored translation models and syntactic translation models.

Factored translation models can be defined as an extension to phrase-based models where every word is substituted by a vector of factors such as word, lemma, part-of-speech information, morphology, etc. The raw training data i.e., the bilingual corpus without factored information cannot be used to generate a factored translation model. Hence, the bilingual corpus has to be factored so that each word in the sentence gets annotated with all the required factors. Till now annotating the factors for Tamil sentences is done manually. Even though various factor generators are available for English, here we coded a factor annotator that uses Stanford parser and a technique to handle phrasal verbs and idioms has been incorporated with it.

The factors for English sentence can be annotated by using an factor annotator to get the necessary information from the Stanford parser in the required format. Also, the same can be done for Tamil but instead of Stanford parser, shallow parser for Tamil has to be used. This will greatly reduce the human effort in annotating the English and Tamil corpus of large size with factors such as word, lemma, part-of-speech information, morphology, etc.

54.3.1 Factor Annotator for English

The factor annotator for English has been coded such that it uses Stanford parser to annotate the factors such as lemma, part-of-speech information, morphology, etc. Here the technique to handle the phrasal verbs and idioms discussed in the previous section has been incorporated with this factor annotator. Reordering module to reorder the word pattern from SVO to SOV has also been incorporated. The block-diagram of the English factor annotator is shown in Figure 4.7.

Fig. 4.7 Block diagram of factor annotator for English

Table 4.6 shows how the factor annotator for English annotates the given English sentence with factors.

Table 4.6 Sample output of factor annotator for English

Input to English Factor Annotator	they are playing .
Output from English Factor Annotator	they|they|PRP|nsubj playing|play|VBG_they_are|root are|be|VBP|aux .|.|.|.

54.3.2 Factor Annotator for Tamil

The factor annotator for Tamil has been coded such that it uses Shallow parser for Tamil to annotate the factors such as lemma, part-of-speech information and morphology. This factor annotator has greatly reduced the human effort employed in annotating the Tamil corpus with factors. The block-diagram of the Tamil factor annotator is shown in Figure 4.8.

Fig. 4.8 Block diagram of factor annotator for Tamil

Table 4.7 shows how the wrapper for English annotates the given English sentence with factors.

Table 4.7 Sample output of factor annotator for Tamil

Input to Tamil Factor Annotator	நான் அவனுக்கு புத்தகத்தைக் கொடுத்தேன்
Output from Tamil	நான் ~PRP~ நான் ~sg அவனுக்கு ~PRP~ அவன் ~sg +dat புத்தகத்தைதக் ~NN~ புத்தகம் ~sg +acc கொடுத்தேன் ~VM~ கொடு~1smf+PAST. ~SYM~&dot~

5.12 Beyond Standard Statistical Machine Translation

Phrase-based models do not consider linguistic information other than words. This linguistic information should be considered, as with this information the quality of the translation would improve. This suggestion leads to the idea of using syntactic information as pre- or post-process e.g. for reordering or re-ranking. There came into picture of the models that include linguistic information in the model itself. They are factored translation models and syntactic-based translation models.

5.12.1 Factored Translation Models

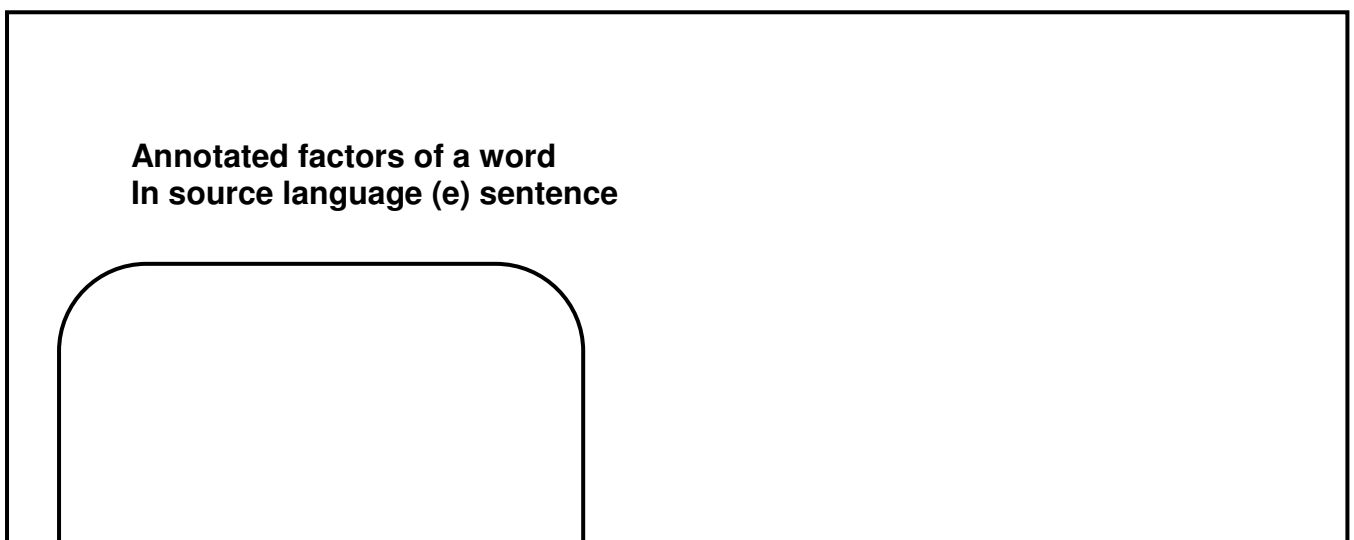
Factored translation models can be defined as an extension to phrase-based models where every word is substituted by a vector of factors such as word, lemma, part-of-speech information, morphology, etc. Here, the translation process has now become a combination of pure translation and generation steps. Figure 5.11 provides a simple block diagram to illustrate the work of translation and generation steps Factored translation models differ from the standard phrase based models from the following:

- The parallel corpus must be annotated with factors such as lemma, part-of-speech, morphology, etc., before training.

- Additional language models for every factor annotated can be used in training the system.

- Translation steps will be similar to standard phrase based systems. But generation steps imply training only on the target side of the corpus.

- Models corresponding to the different factors and components are combined in a log-linear fashion.



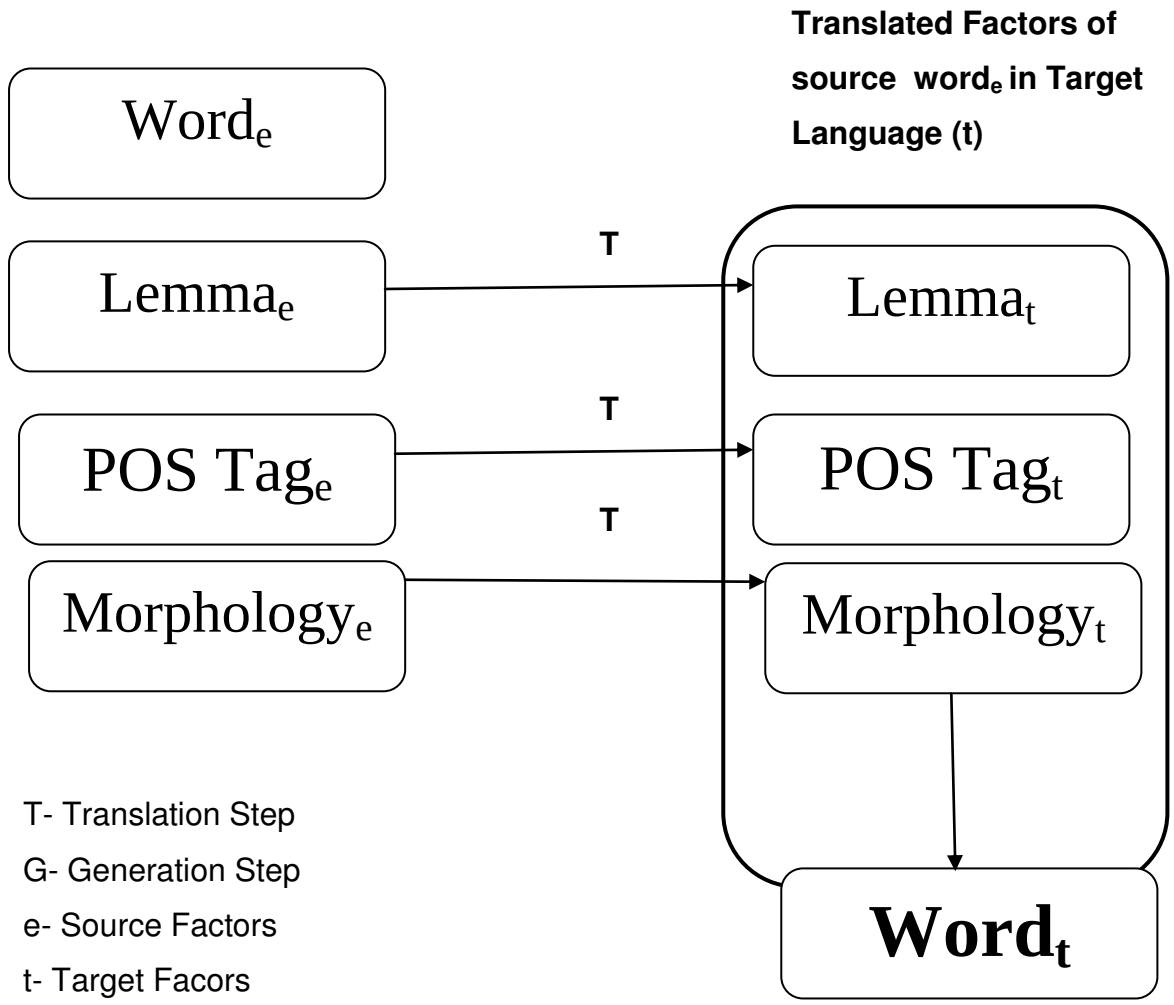


Fig. 5.11 Block diagram to illustrate the work of translation and generation steps

5.12.2 Syntax based Translation Models

Syntax-based translation models use parse-tree representations of the sentences in the training data to learn, among other things, tree transformation probabilities. These methods require a parser for the target language and, in some cases, the source language too. Yamada and Knight propose a model that transforms target language parse trees to source language strings by applying reordering, insertion, and translation operations at each node of the tree. In general, this model incorporates syntax to the source and/or target languages.

Graehl et al. and Melamed**,** propose methods based on tree to tree mappings. Imamura et al. (2005) present a similar method that achieves significant

improvements over a phrase based baseline model for Japanese-English translation. Recently, various preprocessing approaches have been proposed for handling syntax within Statistical machine translation. These algorithms attempt to reconcile the word order differences between the source and target language sentences by reordering the source language data prior to the SMT training and decoding cycles.

Approaches in syntax based models

- Syntactic phrase-based based on tree transducers:
    - Tree-to-string: Build mappings from target parse trees to source strings.
    - String-to-tree: Build mappings from target strings to source parse trees.
    - Tree-to-tree: Mappings from parse trees to parse trees.
- Synchronous grammar formalism that learns grammar can simultaneously generate both trees.
    - Syntax-based: Respect linguistic units in translation.
    - Hierarchical phrase-based: Respect phrases in translation.

5.12.3 On-going Research

Some components of the standard phrase-based model are still under research such as automatic alignments, language models and smoothing techniques and parameter optimization.

Research in adding techniques to improve a standard system such as combining reordering as a pre-process or post-process in the existing system, re-ranking the n-best lists, handling the out of vocabulary words, handling phrasal verbs and idioms, and adapting various domains.

5.13 Summary

Machine English to Tamil Machine Translation System by using parallel Corpus is a novel effort of its kind. The accuracy of the system depends on the amount of parallel corpus available in the languages. Addition of linguistic materials such as morphological information, POS categorization, etc can enhance the accuracy of the system. This is called factored method. At present the system is only in its rudimentary stage. It can translate only simple sentences. Augmentation by increasing the corpus size and inputting linguistic information can enhance the system.

Chapter 6
Conclusion

The present research entitled “English Tamil machine translation system by using parallel corpus” in a novel attempt in the area of machine translation from English to Tamil. Chapter 1 discusses about the thesis work briefly giving the details about the aims and objectives, hypotheses, methodology, earlier works, and use of the research work.

The second chapter makes a brief survey of the MT. Machine Translation (MT) can be defined as the use of computers to automate some or all of the process of translating from one language to another. MT is an area of applied research that draws ideas and techniques from linguistics, computer science, Artificial Intelligence (AI), translation theory, and statistics. Work began in this field as early as in the late 1940s, and various approaches — some ad hoc, others based on elaborate theories — have been tried over the past five decades. This report discusses the statistical approach to MT, which was first suggested by Warren Weaver in 1949 [Weaver, 1949], but has found practical relevance only in the last decade or so. This approach has been made feasible by the vast advances in computer technology, in terms of speed and storage capacity, and the availability of large quantities of text data.

The third chapter discusses about the creation of parallel corpus for building English-Tamil SMT system. Creation of parallel corpus is crucial for building an SMT system. English and other European languages have huge parallel corpora which can be manipulated for SMT. But such parallel corpora are only minimally available for English and Tamil. In the initial stage of SMT we cannot make use of all the available parallel corpora. We have to start with simple sentences and then move on

to complex sentences gradually. Lot of efforts is needed to make the machine to understand the structural differences between these two languages. Sometimes this will be a vexing exercise. One has to keep in mind that the system cannot handle all types of parallel corpora. We have to move very cautiously to get the needed result. We cannot make any tall claim. This chapter is concerned with the creation of parallel corpus for SMT. This chapter discusses about the corpus in general and parallel corpus in particular. The crucial problem in building SMT system is alignment of the corpus. The alignment has to be done in sentence level, phrase level, clause level and word level. All these exercises are time consuming. One should have enough patience to make the computer understand the drastically different two systems of the languages so that it can start translating English into Tamil. Statistical machine translation is one of the alternative methods and not the answer for MT.

One may think that building such system does not require the linguistic knowledge. It is a misnomer. Only if we have full-fledged linguistic knowledge of source language (here English) and Target language (here Tamil) we can attempt to build SMT based system. We cannot build a system simply having the parallel corpus of English and Tamil. Through understanding of the structures of these languages are prerequisite before venture into building such system. This has been done in the 4th chapter. The 4th chapter throws light on how the structures of English and Tamil are different. The difference in structures makes the alignment of these two languages difficult. Factored model makes use of morphological and POS information too.

The fifth chapter discusses about the English to Tamil Machine Translation System by using parallel Corpus. Machine Translation (MT) refers to the use of computers for the task of translating automatically from one language to another. The differences between languages and especially the inherent ambiguity of language make MT a very difficult problem. Traditional approaches to MT have relied on humans supplying linguistic knowledge in the form of rules to transform text in one language to another. Given the vastness of language, this is a highly knowledge intensive task. Statistical MT is a radically different approach that automatically acquires knowledge from large amounts of training data. This knowledge, which is typically in the form of probabilities of various language features, is used to guide the translation process.

Statistical machine translation (SMT) treats the translation of natural language as a machine learning problem. By examining many samples of human-produced translation, SMT algorithms automatically learn how to translate. SMT has made tremendous strides in less than two decades, and new ideas are constantly introduced.

One of the reasons for taking up the SMT approach is ambiguity. Word sense ambiguity and structural ambiguity create great amount of problem in building rule based MT systems. Words and phrases in one language often map to multiple words in another language. For example, in the sentence,

I went to the bank,

it is not clear whether the “mound of sand” (karai in Tamil) sense or the “financial institution” (vangki in Tamil) sense is being used. This will usually be clear from the context, but this kind of disambiguation is generally non-trivial [Nancy and Veronis, 1998]. Also, English and Tamil have their own idiomatic usages which are difficult to identifyfrom a sentence. For example,

India and Pakistan have broken the ice finally.

Phrasal verbs are another feature that are difficult to handle during translation. Consider the use of the phrasal verb bring up in the following sentences,

They brought up the child in luxury. (vaLattu in Tamil)

They brought up the table to the first floor. (meelee koNTu vaa in Tamil)

They brought up the issue in the house. (kavanattiRku koNTuvaa in Tamil)

Yet another kind of ambiguity that is possible is structural ambiguity:

Flying planes can be dangerous.

This can be translated in Tamil as either of the following two sentences.

Vimaanam ooTTuvatu apaayamaanatu

paRakku vimaanangkaL apaayamaanatu

depending on whether it is the planes that are dangerous or the occupation of flying them that is dangerous!

Another reason for undertaking SMT based MT system is structural differences (as we noted in chapter 4) between English and Tamil. Structural Differences English follows a Subject-Verb-Object (SVO) ordering in sentences. Tamil follows SubjectObject-Verb word-ordering in sentences. Apart from this basic

feature, English and Tamil differ also in the structural (or syntactic) constructions that they allow and disallow. These differences have to be respected during translation. For instance, post-modifiers in English become pre-modifiers in Tamil, as can be seen from the following pair of sentences. These sentences also illustrate the SVO and SOV sentence structure in these languages. Here, S is the subject of the sentence, S_m is the subject modifier, (V) is the verb (O) is the object and (O_m is the object modifier.

The president of America will visit the capital of Tamilnadu.

(S) (S_m) (V) (O) (O_m)

Ameerikkak kuTiyaracut talaivar tamiznaaTTin talainakarattiai cenRupaarttaar

(S_m) (S) (O_m) (O) (V)

The structural difference between English and Tamil is discussed elaborately in chapter 4.

Yet another reason for taking up the SMT approach is vocabulary difference. Languages differ in the way they lexically divide the conceptual space, and sometimes no direct equivalent can be found for a particular word or phrase of one language in another. Consider the sentence,

Tendulkar has edged the ball.

edge as a verb has no equivalent in Tamil in this context, and this sentence has to be translated as,

TeTulkar maTTaiyin viLimpu koNTu pantai aTittaar.

Hutchins and Somers (1992] have discussed elaborately about the examples of vocabulary differences between languages and also other problems in MT.

Traditional MT techniques require large amounts of linguistic knowledge to be encoded as rules. Statistical MT provides a way of automatically finding correlations between the features of two languages from a parallel corpus, overcoming to some extent the knowledge bottleneck in MT

Some of statistical models entirely devoid of linguistic knowledge, but similar (non-linguistic) models have achieved encouraging results. Researchers believe that introducing linguistic knowledge can further strengthen the statistical model. Such knowledge may be in the form of morphological rules, rules about word-order, idiomatic usages, known word correspondences and so on. Intuitively, for translation between English and Tamil (or any other Indian language) such linguistic knowledge

might be crucial because of the vast structural and lexical differences between the two languages.

A major drawback with the statistical model is that it presupposes the existence of a sentence-aligned parallel corpus. For the translation model to work well, the corpus has to be large enough that the model can derive reliable probabilities from it, and representative enough of the domain or sub-domain (weather forecasts, match reports, etc.) it is intended to work for. Another issue is that most evaluation of statistical MT has been with training documents that are very rigid translations of each other (parliamentary proceedings have been widely used). News articles and books, for example, are generally rather loosely translated — one sentence in the source language is often split into multiple sentences, multiple sentences are clubbed into one, and the same idea is conveyed in words that are not really exact translations of each other. In such situations, sentence-alignment itself might be a big challenge, let alone word-alignment.

Statistical MT techniques have not so far been widely explored for Indian languages. It would be interesting to find out to what extent these models can contribute to the huge ongoing MT efforts in the country.

Since statistical MT is in some sense word alignment (with probabilities), it can be used for lexicon acquisition also, apart from the larger goal of MT. The present system is only in its initial stage. Augmentation by increasing the corpus size and adding linguistic information can enhance the accuracy of the system.

Appendix 1: A sample of English and Tamil parallel Corpus

Sno	**English Corpus**	**Tamil Corpus**
1	When the plasmodium parasite enters your bloodstream, it travels to the liver and then re-enters the blood stream where it is able to invade red blood cells.	பிளாஸ்மோடியம் ஒட்டுண்ணி உங்கள் இரத்த ஓட்டத்தில் நுழையும் போது, அது கல்லீரலுக்கு பயணிக்கிறது; பின்னர் அது சிவப்பு இரத்த அணுக்களைத் தாக்கவியலும் இரத்த ஓட்டத்தில் மீண்டும் நுழைகிறது.
2	Eventually, the infected red blood cells burst and, when they do, they release even more of the tiny parasites into the blood.	இறுதியாக, தொற்றப்பட்ட சிவப்பு இரத்த அணுக்கள் வெடிக்கும், அவ்வாறு செய்கையில் அவை இரத்தில் மேலும் கூடுதலான சிறிய ஒட்டுண்ணிகளை

		வெளியீடும்.
3	The infected cells tend to burst every 48-72 hours.	பாதிக்கப்பட்ட செல்கள், ஒவ்வொரு 48-72 மணிக்கும் வெடிக்க முனைகின்றன.
4	Each time they burst, you will usually experience an attack of chills, fever and sweating.	அவை வெடிக்கும் ஒவ்வொரு முறையும் நீங்கள் பொதுவாகக் குளிரின் தாக்கம், காய்ச்சல் மற்றும் வியர்த்தல் இவற்றை அனுபவிப்பீர்கள்.
5	Research suggests that if you are pregnant, you are more at risk of catching malaria than normal.	நீங்கள் கர்ப்பமாக இருந்தால், சாதாரணர்களை விட உங்களை மலேரியா தொற்றும் அபாயம் அதிகம் என்று ஆராய்ச்சி கூறுகிறது.
6	This is because your immune system can be weaker during pregnancy, which means your body is less able to fight off bacteria and infection.	ஏனென்றால் இது உங்கள் நோயெதிர்ப்பு கர்ப்பத்தின் போது பலவீனமான முடியும் , அது உங்கள் உடலில் பாக்டீரியா மற்றும் தொற்றுடன் போராட குறைவாகவே முடியும் என்றாகிறது.
7	If you are pregnant and have malaria, you may pass the infection on to your baby.	நீங்கள் கர்ப்பமாக இருந்து மலேரியாவும் இருந்தால் நீங்கள் உங்கள் குழந்தைக்கு இந்நோயைக் கடத்தக் கூடும்.
8	See the 'treatment' section for details of the malaria medicines that are safe to use during pregnancy.	கர்ப்பத்தின் போது பயன்படுத்தப் பாதுகாப்பான மலேரியா மருந்துகளின் விவரங்களுக்கு 'சிகிச்சை' என்ற பகுதியைப் பார்க்கவும்.
9	Like most viral infections, the chickenpox virus is transmitted from person to person by droplet spread, e. g. , sneezing, and by close contact with an infected person such as touching the fluid oozing from the blisters or using the infected person's clothes or towels.	பெரும்பாலான வைரஸ் தொற்று நோய்களைப் போலவே சின்னம்மை வைரசும் நபருக்கு நபர் சிறு துளியின் பரவலால் கடத்தப்படுகின்றது, எ.கா. , தும்மல் மற்றும் கொப்புளங்களிலிருந்து கசிகிற திரவத்தைத் தொடுதல் போன்ற நெருக்கமான தொடர்பாலோ அல்லது பாதிக்கப்பட்ட நபரின் துணிகளையோ துவாலைகளையோ பயன்படுத்துவதாலோ.
10	Most children will contract chickenpox when they are young and it is usually a mild illness in children, though can be more serious in teenagers	அவர்கள் இளவயதின் போது பெரும்பாலான குழந்தைகள் சின்னம்மை ஒப்பந்தத்தையும் மற்றும் இளம் பருவத்தினர் மற்றும் வயது வந்தோருக்கு

	and adults.	தீவிரமான இருக்கலாம் என்றாலும், பொதுவாக குழந்தைகளுக்கு லேசான உடல்நல குறைவு உள்ளது.
11	The incubation period, i. e. , the time taken from when the disease is first caught until the symptoms appear, is from 14 - 21 days.	அடைகாக்கும் காலம், அதாவது முதலில் நோய் தொற்றியதிலிருந்து அறிகுறிகள் தோன்றும் வரை எடுத்துக்கொள்ளப்பட்ட காலம் 14-இலிருந்து 21 நாட்களாகும்.
12	The child is infectious from about two days before the rash appears until all the spots have dried up which can take up to ten days.	தோல் வெடிப்பு தோன்றுவதற்கு முன்னர் கிட்டத்தட்ட இரண்டு நாட்களிலிருந்து அனைத்துப் புள்ளிகளும் உலர்கிறது வரை குழந்தை தொற்றுவிப்பதாய் இருக்கும்; இது பத்து நாட்கள் வரை எடுக்கும்.
13	Most people get chickenpox at least once in their lifetime.	பெரும்பாலான மக்கள் தங்கள் வாழ்நாளில் குறைந்தது ஒருமுறையாவது சின்னம்மை நோயைப் பெறுகின்றனர்.
14	If you have had chicken pox before, it is very unlikely that you would get it again.	உங்களுக்கு சின்னம்மை இருந்து இருந்தால், அது உங்களுக்கு மீண்டும் கிடைக்கும் வாய்ப்பு மிகவும் குறைவாக உள்ளது.
15	The infection produces antibodies that help fight the virus, if it attacks again.	வைரஸின் மறுபடியும் தாக்குதல்கள் இருந்தால் தொற்று போராட உதவும் பிறபொருளெதிரிகளை அது உருவாக்குகிறது.
16	It is more common among children below ten years.	இது பத்து ஆண்டுகளுக்கு கீழே குழந்தைகளுக்கு மிகவும் பொதுவானதாக உள்ளது.
17	The infection is mild in children but the severity increases in adults and pregnant women.	நோய்த்தொற்று குழந்தைகளுக்கு லேசாக உள்ளது ஆனால் தீவிரத்தன்மை பெரியவர்களில் மற்றும் கர்ப்பிணி பெண்களுக்கு அதிகரிக்கிறது.
18	The incidence of chickenpox is the highest during summers.	சின்னம்மை நோயின் நிகழ்வு கோடை காலத்தில் அதிகமாக உள்ளது.

19	The flu virus family contains three main types, A, B and C. Flu viruses are known to mutate (change) over time.	ஃப்ளூ வைரஸ் குடும்பத்தில் ஏ, பி, சி என்ற மூன்று முக்கிய வகைகள் உள்ளன. ஃப்ளூ வைரஸ்கள் காலத்தால் மாறக்கூடியன என அறியப்படுகின்றன.
20	Also, strains of the flu virus are classified on where and when they were first identified and given a serial number.	மேலும், ஃப்ளூ வைரசின் சந்ததிகள் எங்கு, எப்போது அவை முதலில் அடையாளம் காணப்பட்டன என வகைப்படுத்தப்பட்டு வரிசை எண் தரப்பட்டுள்ளன.
21	Flu A occurs more frequently (every two to three years) and is more serious than type B. It is very likely to mutate and it regularly produces varieties to which populations have no resistance.	ஃப்ளூ ஏ பி-ஐ விட அடிக்கடி ஏற்படுகிறது (ஒவ்வொரு இரண்டு அல்லது மூன்று ஆண்டுகள்) மற்றும் பி வகையை விட மிகக் கடுமையானது. இது அடிக்கடி மாறக்கூடியது மற்றும் இது வழக்கமாகப் பல வகைகளை உற்பத்தி செய்கின்றது; இவற்றிற்கு மக்களுக்கு எதிர்ப்புசக்தி இல்லை.
22	It is for this reason that widespread epidemics occur that may affect whole continents.	இக்காரணத்தால் தான் முழு கண்டத்தையே பாதிக்கும் பரவலான கொள்ளைநோய் ஏற்படுகின்றது.
23	These are known as pandemics and are caused by new strains of the type A virus.	இவை உலகளாவியத் தொற்றுநோய் என்று அறியப்படுகின்றது; மற்றும் இவை A வகை வைரஸின் புதிய சந்ததிகளால் ஏற்படுகின்றன.
24	Generally, flu B causes a less severe illness, although it is responsible for smaller outbreaks.	சிறிய திடீர் நோய் தாக்கத்திற்குக் இது காரணமாக இருந்தாலும், பொதுவாக ஃப்ளூ பி குறைந்த அளவு கடுமையான உடல்நலக்குறைவையே ஏற்படுத்துகிறது.
25	Flu B is much more stable than the flu A virus and if you have been infected with flu B, your immunity to further flu B infections will last for many years.	ஃப்ளூ பி ஃபுளூ ஏ வைரசை விட கூடுதல் நிரந்தரமானது; மற்றும் நீங்கள் ஃப்ளூ பி-ஆல் தொற்றப்பட்டிருந்தால் திரும்பவரும் ஃப்ளூ பி தொற்றுக்களுக்கு உங்கள் எதிர்ப்பு சக்தி பல ஆண்டுகளுக்கு நீடிக்கும்.
26	Flu B mainly affects young children (5-14 years of age) who have not been exposed to the virus and have not developed immunity.	ஃப்ளூ பி முக்கியமாக வைரசுக்கு ஆளாக்கப்படாத மற்றும் நோய் எதிர்ப்பு சக்தி உருவாக்கிக்கொள்ளாத இளம் குழந்தைகளை (5-14வயது ஆண்டுகள்) பாதிக்கிறது.

27	In the winter of 2005/6, the majority of flu activity was confined to type B with only a few cases of flu A reported.	2005/6 குளிர்காலத்தில், ஃப்ளூ ஏ-இன் ஒரு சில நிகழ்வுகள் மட்டுமே தெரிவிக்கப்பட்டதால் பெரும்பான்மையான ஃப்ளூ செயல்பாடு பி வகைக்கு மட்டுமே வரையறுக்கப்பட்டது.
28	Type C usually causes a mild illness similar to the common cold.	சி வகை பொதுவாக சாதாரண நீர்கோப்பு போன்ற மிதமான உடல்நலக்குறைவையே ஏற்படுத்தும்.
29	In recent years, two subtypes of flu A have been circulating, the H1N1 subtype and the H3N2 subtype - Panama or Moscow-like strains.	அண்மை ஆண்டுகளில் பனாமா அல்லது மாஸ்கோ போன்ற சந்ததிகளான ஃப்ளூ ஏ-இன் இரண்டு துணைவகைகளான எச்1என்1 துணைவகையும் எச்3என்2 துணைவகையும் பரவிவருகிறது.
30	In 2003/4, the main strain circulating was a flu A (H3N2) Fujian-like strain.	2003/4-இல் பரவிக்கொண்டிருந்த முக்கிய சந்ததி ஃபுஜியன் சந்ததி போன்ற ஃப்ளூ ஏ (எச்3என்2) ஆகும்.
31	This is slightly different to the A (H3N2) Panama-like virus, which has been circulating in the UK in recent years.	இது அண்மை ஆண்டுகளில் யுகேயில் பரவிவருகிற பனாமா போன்ற வைரஸ் ஏ (எச்3என்2)-இலிருந்து சற்று வேறுபட்டது.
32	The hepatitis A virus is in the stools (faeces) of affected people.	ஹெப்படைடிஸ் ஏ வைரஸ் பாதிக்கப்பட்ட மக்களின் மலத்தில் உள்ளது.
33	The disease is easily spread in areas where there is overcrowding and poor sanitation.	மக்கள் நெருக்கமும் மோசமான சுகாதார வசதியும் உள்ள பகுதிகளில் இந்நோய் எளிதில் பரவுகிறது.
34	The most common cause of infection with the hepatitis A virus is via the faecal-oral route, which passes the virus on from person to person due to poor personal hygiene.	ஹெப்படைடிஸ் A வைரஸ் தொற்றவதற்கு மிகப் பொதுவான காரணம் மலம் வழியும் வாய்வழியும் ஆகும்; இது மோசமான தனிப்பட்ட சுகாதாரம் காரணமாக நபருக்கு நபர் வைரசைக் கடத்துகின்றது.
35	For example, you may get hepatitis A if you eat food prepared by an infected person who has not properly washed their hands.	எடுத்துக்காட்டாக, நீங்கள் தனது கைகளைச் சரியாகக் கழுவாதிருந்திருக்கிற நோய் தொற்றிய ஒரு நபரால் உண்டாக்கப்பட்ட உணவை உண்டால் ஹெபடைடிஸ் ஏ-ஐப் பெறக்கூடும்;

36	It is also possible to become infected by drinking water, which has become contaminated due to inadequate sewage treatment.	இது போதாத கழிவுநீர் சுத்திகரிப்பு காரணமாக அசுத்தமான மாறியுள்ள குடிநீர், மூலம் தொற்று சாத்தியமுள்ளதாக இருக்கிறது.
37	Uncooked foods, including raw fruit and vegetables, untreated drinking water, and ice cubes, and food prepared or washed with contaminated water, can all transmit the viral infection.	வேகவைக்காத பழம் மற்றும் காய்கறிகள், சுத்திகரிக்கப்படாத குடிநீர், மற்றும் ஐஸ் க்யூப்ஸ், மற்றும் அசுத்தமான தண்ணீர் உணவு தயாரித்த அல்லது கழுவப்பட்ட உணவுகள் உட்பட, அனைத்து வைரஸ் கிருமி தொற்று பரவ முடிகிறது.
38	Also, shellfish can be infected if it comes from sea that is contaminated with sewage.	மேலும், சிப்பி மீன் சாக்கடையால் மாசுபடுத்தப்பட்ட கடலில் இருந்து வந்தால் அது நோயால் தொற்றப்பட்டிருக்கலாம்.
39	Hepatitis B is spread when blood or body fluids from someone who has the virus infects someone who is not immune.	வைரஸ் உள்ள ஒருவரிடமிருந்து இரத்தமோ உடல் திரவங்களோ எதிர்ப்பு சக்தி இல்லாத மற்றொருவரை தொற்றும் போது ஹெபடைடிஸ் பி பரப்பப்படும்.
40	Many people with hepatitis B do not even realise that they are infected.	ஹெபடைடிஸ் பி உள்ள பல மக்கள் தாங்கள் நோயால் தொற்றப்பட்டவர்கள் என்று கூட அறிவதில்லை.
41	There is a small risk of contracting the hepatitis B virus from sharing toothbrushes, razors and towels, which may be contaminated with blood.	இரத்தத்தால் மாசுபடுத்தப்பட்டு இருக்கவியலும் டூத்ப்ரஷ்கள், ரேசர்கள் மற்றும் துவாலைகள் இவற்றைப் பகிர்ந்துகொள்ளுதல் மூலம் ஹெபடைடிஸ் பி வைரஸ் பரவும் ஒரு சிறிய ஆபத்து இருக்கிறது.
42	Unsafe tattooing and body piercing practices also risk potentially spreading the virus.	பாதுகாப்பற்ற பச்சைக்குத்துதலும் உடலில் துளையிடும் நடைமுறைகளும் வைரஸ் பரவும் வாய்ப்புக்கான அபாயம் உள்ளது.
43	The viral infection is also more likely to be passed on in countries where equipment for medical and dental treatment is not sterilised properly.	மேலும் மருத்துவம் மற்றும் சிகிச்சைக்கான கருவிகளில் நோய்க் கிருமிகள் சரியாக அழிக்கப்படாத நாடுகளில் வைரல் தொற்றுநோய் பரவும் வாய்ப்பு அதிகம் உள்ளது.
44	This can also be the case in countries where blood is not tested for hepatitis B, and blood transfusions may still result in infection.	ஹெபடைடிஸ் பி-க்கு வேண்டி இரத்தம் பரிசோதனை செய்யப்படாத நாடுகளிலும் இது நிகழும்; மற்றும் இரத்தம் செலுத்துதல் இன்னும் தொற்று நோய் பீடிப்பில் முடிவுறலாம்.

45	All blood donations in the UK are tested for hepatitis B. Travellers are advised to vaccinate themselves against hepatitis B before setting off on their trip.	யுகேயில் அனைத்து இரத்த தானங்களும் ஹெபடைடிஸ் பி-க்கு வேண்டி பரிசோதனை செய்யப்படுகின்றன. பயணிகள் தங்கள் பயணத்தைத் தொடங்குவதற்கு முன் ஹெபடைடிஸ் பி-க்கு எதிராகத் தங்களுக்கு நோய்தடுப்பு ஊசி போட்டுக்கொள்ள அறிவுரை செய்யப்படுகின்றனர்.
46	Chronic hepatitis (persistent liver inflammation) can also be caused by the body attacking its own organs as if they were a foreign bacteria or infection.	ஒரு அந்நிய பாக்டீரியா அல்லது தொற்று நோய் பீடிப்பு இருப்பதாக எடுத்துக்கொண்டு தன் சொந்த உறுப்புகளையே உடல் தாக்குவதாலும் நாள்பட்ட ஹெபடைடிஸ் (தொடர்ந்திருக்கிற கல்லீரல் அழற்சி) ஏற்படலாம்.
47	This is known as autoimmune hepatitis and is a rare cause of chronic hepatitis.	இது தன் தடுப்பாற்று ஹெப்படைடிஸ் என்று அழைக்கப்படுகிறது; மற்றும் நாள்பட்ட ஹெபடைடிஸ் ஒரு அரிதான காரணம் ஆகும்.
48	Hepatitis B cannot be spread through sneezing, coughing or hugging someone who is infected with the viral infection.	ஹெபடைடிஸ் பி-ஐ தும்மல், இருமல் அல்லது வைரல் நோயால் பீடிக்கப்பட்ட ஒருவரைக் கட்டியணைத்தல் மூலம் பரப்ப இயலாது.
49	Measles is caused by infection with the rubeola virus.	தட்டம்மை ருபீயோலா வைரசின் பீடிப்பால் ஏற்படுகின்றது.
50	Once infected, the virus lives in the mucus of the nose and throat.	ஒருமுறை நோய் தொற்றிக்கொண்டால் இந்த வைரஸ் மூக்கு மற்றும் தொண்டை சளியில் வாழும்.
51	Physical contact, coughing and sneezing can spread the infection.	உடல் தொடர்பு, இருமல் மற்றும் தும்மல் என்பன தொற்று நோயைப் பரப்ப இயலும்.
52	Infected droplets of mucus may also land on a surface where they remain active and contagious for around two hours.	நோய் பீடிக்கப்பட்ட சளியின் திவலைகள் மேற்பரப்பில் விழலாம்; அங்கு நோய் கிருமிகள் சுமார் இரண்டு மணி நேரம் செயலூக்கத்துடனும் தொற்றும் நிலையிலும் இருக்கும்.
53	Once inside your body, the virus multiplies in the back of your throat and lungs, before spreading throughout your body, including your	ஒருமுறை உங்கள் உடலின் உள்ளே நுழைந்துவிட்டால், உங்கள் சுவாச ஒழுங்கமைப்பு மற்றும் தோல் உட்பட உங்கள் உடல் முழுவதும்

	respiratory system and the skin.	பரவும் முன், இந்த வைரஸ் உங்கள் தொண்டை மற்றும் நுரையீரலின் பின்னால் பன்மடங்காகப் பெருகும்.
54	It takes between 6-21 days for the virus to establish itself (the incubation period), but people usually show symptoms after about 10 days.	இந்த வைரஸ் தன்னை நிலைநிறுத்திக்கொள்ள 6-21 நாட்கள் (நோய்காப்பு காலம்) எடுத்துக்கொள்ளும். ஆனால் மக்கள் பொதுவாகச் சுமார் 10 நாட்களுக்குப் பிறகு நோய்க்கான அறிகுறியைக் காட்டுவர்.
55	Someone with measles is infectious for 2 to 4 days before the red rash appears and for about five days after it appears.	தட்டம்மை உள்ள ஒருவர் சிவப்பு வெடிப்பு தோன்றும் முன் 2 முதல் 4 நாட்களுக்கும் தோன்றிய பிறகு சுமார் ஐந்து நாட்களுக்கும் தொற்று நோயைப் பரப்பக்கூடியவராக இருப்பர்.
56	Anyone who has not had measles before can be infected.	முன்பு தட்டம்மை வராதவர் எவரையும் நோய் தொற்ற இயலும்.
57	However, cases of re-infection after having had the virus are extremely rare because the body will have built up immunity to the virus.	எனினும், வைரசால் பாதிக்கப்பட்ட பின்னர் மீண்டும் தொற்றும் நிகழ்வுகள் மிக அரிதாகும்; ஏனென்றால் வைரஸ் தடுப்பாற்றலை உடல் உருவாக்கிக்கொள்ளும்.
58	About 90% of people, who are not immune from measles and are sharing a house with somebody who is infected, will develop the condition.	தட்டம்மை எதிர்ப்பு சக்தி இல்லாதவர்கள் நோய் தொற்றப்பட்ட ஒருவருடன் ஒரு வீட்டைப் பகிர்ந்து கொண்டிருந்தால் சுமார் 90% மக்கள் அந்நோயால் பாதிக்கப்படுவர்.
59	Measles virus belongs to the Morbillivirus group of the Paramyxovirus family.	தட்டம்மை வைரஸ் பாராமிக்ஸோவைரஸ் குடும்பத்தின் மார்பிலிவைரஸ் குழுக்குள் அடங்குகிறது.
60	Humans are the only natural host for wild measles virus.	மனிதர்கள் மட்டுமே முரட்டுத் தட்டம்மை வைரசின் இயல்பான ஆதார உயிரியாக உள்ளனர்.
61	The virus is easily destroyed but remains in the droplet form in air for several hours, especially under conditions of low relative humidity.	வைரஸ் எளிதில் அழிக்கப்பட்டுவிடும், ஆனால் முக்கியமாகக் குறைந்த ஒப்பு ஈரப்பதச் சூழ்நிலைகளில் பல மணி நேரம் காற்றில் திவலை வடிவில் எஞ்சியிருக்கும்.

62	It is spread by direct contact with droplets from respiratory secretions of infected persons.	இது தொற்று ஏற்பட்டவர்களில் சுவாசத்திலிருந்து வெளிவரும் சிறுதுளிகளின் நேரடி தொடர்பு மூலம் பரவுகிறது.
63	It is one of the most communicable of infectious diseases and is most infectious when cough and cold is at its peak.	இது தொற்று நோய்களுள் மிக அதிகமாகப் பரவக்கூடிய ஒன்றாக உள்ளது; மற்றும் இருமலும் தடுமனும் உச்சத்தில் இருக்கும் போது மிக அதிகமாகத் தொற்றக்கூடியதாக இருக்கும்.
64	The virus invades the respiratory lining membrane and then enters the blood stream.	இவ்வைரஸ் சுவாச உட்புற மென்படலத்தைத் தாக்கும்; பின்னர் இரத்த ஓட்டத்தில் நுழையும்.
65	It causes inflammation of the respiratory tract and may predispose to secondary bacterial pneumonia.	இது சுவாசக்குழாய் அழற்சியை உருவாக்குகிறது மற்றும் இரண்டாம் நிலை பாக்டீரியாசார்ந்த நிமோனியாவுக்குப் பொறுப்புள்ளதாகும்.
66	Malaria is caused due to infection by the protozoan Plasmodium species.	புரோட்டோசோவன் பிளாஸ்மோடியம் இனத்தின் தொற்றுதல் காரணமாக மலேரியா ஏற்படுகிறது.
67	It is transmitted by the bite of the infected Anopheles mosquito.	இது தொற்றப்பட்ட அனாஃபிலிஸ் கொசுக்கள் கடிப்பதன் மூலம் பரவுகிறது.
68	Four major species of Plasmodia are implicated in the causation of malaria in humans and these are Plasmodium Vivax, Plasmodium Ovale, Plasmodium Malariae and Plasmodium Falciparum.	பிளாஸ்மோடியாவின் நான்கு முக்கிய இனங்கள் மனிதர்களுக்கு மலேரியா உருவாகக் காரணமாகச் சுட்டப்படுகிறன; இவை பிளாஸ்மோடியம் விவக்ஸ், பிளாஸ்மோடியம் ஓவலே, பிளாஸ்மோடியம் மலேரியே மற்றும் பிளாஸ்மோடியம் ஃபால்ஸிபரம் என்பனவாகும்.
69	Among these species, it is Plasmodium Falciparum that is the most dangerous and that is responsible for most of the deaths resulting from malaria.	இந்த இனங்களுக்குள் பிளாஸ்மோடியம் ஃபால்ஸிபரம் மிகவும் ஆபத்தானது; மற்றும் இது மலேரியாவினால் ஏற்படும் இறப்புகளுள் பொரும்பான்மையானவைகளுக்குப் பொறுப்பானதாகும்.
70	Infection begins when the infected female Anopheles mosquito bites the human.	நோயால் பீடிக்கப்பட்ட பெண் அனாஃபிலிஸ் கொசு மனித கடிக்கும் போது நோய் பீடிப்பு துவங்குகிறது.

71	The microscopic forms of the parasite are carried through the patient's blood stream until they reach the liver.	ஒட்டுண்ணிகள் நுண்ணிய படிவங்களை நோயாளியின் இரத்த ஓட்டத்தின் வழியே அவை கல்லீரல் அடையும் வரை எடுத்து செல்லப்படுகிறது.
72	There they invade the liver cells and begin to reproduce.	அங்கு அவைகள் கல்லீரல் செல்களைத் தாக்குகின்றன; மற்றும் இனப்பெருக்கம் செய்கின்றன.
73	The swollen liver cells eventually burst discharging the merozoite forms of the parasite into circulation and this is when the symptoms of the infection start to become apparent.	வீங்கிய கல்லீரல் செல்கள் இறுதியில் வெடித்து ஒட்டுண்ணியின் மெரொசோயிட் வடிவுகளை இரத்தச் சுற்றோட்டத்தில் வெளியேற்றுகின்றன; இப்போதுதான் நோய் தொற்றியதன் அறிகுறிகள் வெளிப்படையாகும்.
74	Once inside the bloodstream the merozoites invade the red blood cells and start to grow.	இரத்த ஓட்டத்தின் உள்ளே வந்ததும் மெரொசோயிட்கள் சிவப்பு இரத்த அணுக்களைத் தாக்குகின்றன மற்றும் வளரத் தொடங்குகின்றன.
75	They consume and degrade the intracellular proteins inside the red cell, especially the haemoglobin, eventually causing the infected red cells to rupture.	அவை சிவப்பு அணுக்களை உண்கின்றன மற்றும் சிவப்பு அணுக்களின் உள்ளிருக்கிற செல்லிடை புரதங்களை சீர்கெடச்செய்கிறன. குறிப்பாக ஹீமோகுளோபினைச் சிர்கெடச்செய்கின்றன. இறுதியாகப் பாதிக்கப்பட்ட சிவப்பு அணுக்களைச் சிதையச் செய்கின்றன.
76	Salmonella are non-encapsulated, rod-shaped organisms that are motile by means of flagellae.	சால்மோனெல்லா காப்புறையற்ற கோல் வடிவ உயிரினங்கள் ஆகும்; அவை நகரிழைகள் மூலம் இடம்பெயர்பவை.
77	They express several antigens, including 'H' and 'O'.	அவைகள் 'எச்' மற்றும் 'ஓ' உட்பட்ட பல ஆன்டிஜென்களை வெளிப்படுத்துகின்றன.
78	These bacteria, after ingestion through contaminated food, move into the small bowel where they interact with the intestinal wall.	இந்தப் பாக்டீரியாக்கள் அசுத்தமான உணவு மூலம் உட்கொள்ளப்பட்டதும் அவை சிறு குடலுக்குள் செல்கின்றன; அங்கே சிறு குடல் சுவருடன் எதிர்வினைபுரிகின்றது.
79	After they enter the intestinal wall, they survive in macrophages, which are white blood cells that swallow the typhoid bacilli.	அவை சிறு குடல் சுவரில் நுழைந்த பிறகு அவை டைபாய்ட் பாசில்லியை விழுங்கும் வெள்ளை இரத்த அணுக்களான மாக்ரோபேஜசில் தொடர்ந்துவாழும்.

80	They are then disseminated to several organs through the bloodstream.	அவை பின்னர் இரத்த ஓட்டத்தின் மூலம் பல உறுப்புகளுக்கும் பரவும்.
81	There is a secondary phase where the typhoid bacilli enter the blood stream to cause the clinical symptoms of typhoid.	இரண்டாம் கட்டத்தில் டைபாய்டு பாக்டீரியா இரத்த ஓட்டத்தில் நுழைந்து டைபாய்டின் அறிகுறிகளை ஏற்படுத்தும்.
82	They are also excreted in the urine and faeces.	அவை சிறுநீர் மற்றும் மலத்தின் மூலம் வெளியேற்றப்படும்.
83	The mumps virus is spread in saliva and in minute airborne droplets from the coughs and sneezes of infected people.	தாளம்மை வைரஸ் உமிழ்நீரில் பரவுகின்றது; மற்றும் நோய் தொற்றப்பட்ட மக்களின் இருமல்கள் மற்றும் தும்மல்களிலிருந்து வரும் திவலைகளிலிருந்து நிமிடத்தில் காற்றில் பரவும்.
84	The virus that most commonly causes mumps is a paramyxovirus.	பொதுவாகத் தாளம்மையை உண்டாக்கும் வைரஸ் ஒரு பாராமைக்ஸோவைரஸ் ஆகும்.
85	On rare occasions your parotid glands may swell due to the influenza virus or Coxsackie virus.	அரிதாக உங்கள் பேரோடிட் சுரப்பிகள் இன்ஃப்ளூயன்ஸா வைரஸ் அல்லது கோக்ஸாக்கீ வைரஸ் காரணமாக வீங்கக்கூடும்.
86	These conditions may also be responsible for mumps returning.	இந்த நிலைமைகள் கூட தாளம்மை திரும்பவருவதற்குப் பொறுப்பாக இருக்கலாம்.
87	Lack of immunisation is the main factor that puts people at risk of having mumps.	நோய் எதிர்ப்புசக்தி ஊட்டுவதில் உள்ள குறைபாடு தான் தட்டமை வருகிற ஆபத்தில் மக்களைச் சிக்கவைக்கின்ற முக்கிய காரணியாக இருக்கின்றது.
88	If you were not given a vaccination as a child, or have a weakened immune system, you may also be at risk.	நீங்கள் குழந்தையாக இருக்கும் போது நோய் தடுப்பு ஊசி போடப்படவில்லை என்றாலோ ஒரு நலிவுற்ற நோய் எதிர்ப்பு ஒழுங்கமைப்பு இருந்தாலோ நீங்களும் ஆபத்தில் இருக்கக்கூடும்.
89	Polio is caused by a particular type of virus known as an enterovirus.	போலியோ எண்ட்ரோவைரஸ் எனப்படும் ஒரு குறிப்பிட்ட வகை வைரசால் ஏற்படுகிறது.

90	An enterovirus is a type of virus which grows and thrives in the gastrointestinal tract (the system of organs which help digest food, such as the intestines and stomach).	எண்ட்ரோ வைரஸ் இரைப்பை-குடல் பாதை (குடல் மற்றும் வயிறு போன்ற உணவின் செரிமானத்திற்கு உதவும் உறுப்புகளின் ஒழுங்கமைப்பு) வளர்கிற மற்றும் பெருக்கமடைகிற ஒரு வைரஸ் வகையாகும்.
91	After growing in the gastrointestinal tract, an enterovirus will often move on to affect the nervous system.	இரைப்பை-குடல் பாதையில் வளர்ந்தபிறகு ஒரு எண்ட்ரோ வைரஸ் பிறகு பெரும்பாலும் நரம்பு மண்டலத்தை பாதிக்க முன்னேறிவிடும்.
92	The polio virus is usually spread through the faeces of someone infected with the illness.	போலியோ வைரஸ் பொதுவாக நோய் தொற்றப்பட்ட ஒருவரின் மலத்தின் மூலம் பரவுகிறது.
93	This is why polio tends to be more common in less well developed countries with poor sanitation.	எனவேதான் சுகாதார குறைவு உள்ள நன்கு வளராத நாடுகளில் போலியோ பொதுவாக இருக்கும் நிலை உள்ளது.
94	The virus is most easily transmitted when someone has oral contact with infected faeces.	ஒருவர் நோயால் பீடிக்கபட்ட மலத்தின் வாய் வழி தொடர்பின் போது வைரஸ் மிக எளிதில் பரவுகிறது.
95	This usually involves a person drinking water which has been contaminated with infected faeces.	இது பொதுவாக ஒரு நபர் நோயால் பீடிக்கப்பட மலத்தினால் மாசுபடுத்தப்பட்ட குடிநீரைப் பருகுவதை உட்படுத்துகிறது.
96	Polio can also be spread through contaminated water and food.	போலியோ அசுத்தமான தண்ணீர் மற்றும் உணவு மூலமும் பரவ இயலும்.
97	In some, rare cases, it can also be transmitted through direct contact with someone who is infected.	சில அரிய சந்தர்பங்களில் அது நோயால் பீடிக்கப்பட்ட ஒருவரின் நேரடி தொடர்பு மூலமும் பரவ இயலும்.
98	This is because the virus will be present in the saliva of an infected person so contact, such as kissing may help spread the infection.	இந்த வைரஸ் தொற்று நோயால் பாதிக்கப்பட்ட நபரின் உமிழ்நீரில் இருப்பதால் முத்தம் போன்ற தொடர்பு தொற்று நோயைப் பரப்ப உதவும்.
99	Once the polio virus enters your body, it begins to multiply in you throat and intestines.	போலியோ வைரஸ் உங்கள் உடலில் ஒரு முறை நுழைந்துவிட்டால் அது உங்கள் தொண்டை மற்றும் குடகளில் பெருகத் தொடங்கும்.

100	It then travels to your central nervous system through your blood.	இது பின்னர் உங்கள் இரத்தம் மூலம் உங்கள் மைய நரம்பு மண்டலத்திற்குப் பயணிக்கும்.

BIBLIOGRAPHY

Aarts, J and Meijs, W (eds.) 1984. Corpus linguistics. Rodopi, Amsterdam.

Aarts, J and Meijs,W (eds.) 1986. Corpus linguistics II. Rodopi, Amsterdam.

Aarts, J and Meijs,W (eds.) 1990. Theory and practice in corpus linguistics. Rodopi, Amsterdam.

Ahrenberg, Lars, Mikael Andersson, and Magnus Merkel, 2000. Parallel text processing: Alignment and Use of Translation Corpora, volume 13 of Text, Speech and Language Technology, chapter 5 — “A Knowledge lite approach to word alignment”, pages 97–116. Kluwer Academic Publishers.

Aijmer A and Altenberg B (eds.) 1991. English Corpus Linguistics: Studies in honour of Jan Svartvik, Longman, London.

ALPAC. 1966. Language and Machines: Computers in Translation and Linguistics. A report by the Automatic Language Processing Advisory Committee (Tech. Rep. No. Publication 1416). 2101 Constitution Avenue, Washington D.C., 20418 USA: National Academy of Sciences, National Research Council.

Annamalai, E. “Corpora Development in Indian Languages”, in Agarawal and Pani (eds.) Information Technology Applications in Language, Script and Speech, New Delhi: BPB Publication.

Arnold D. 2003. “Why translation is difficult for computers” in Somers, H. (ed.). 2003. Computers and Translation: A translator’s guide. Benjamins Translation Library.

Bandyopadhyay S. 2000. ANUBAAD - The Translator from English to Indian Languages. In proceedings of the VIIth State Science and Technology Congress. Calcutta. India. pp. 43-51.

Bharati, Akshar, Chaitanya, Vineet, Kulkarni, Amba P., Sangal, Rajeev. 1997. Anusaaraka: Machine Translation in stages. Vivek, A Quarterly in Artificial Intelligence, Vol. 10, No. 3. , NCST, Banglore. India, pp. 22-25.

Bloom, L.1970. Language development: form and function in emerging grammars, MIT press, Cambridge.

Boas, F. 1940. Race, language and culture. Macmillan, New York.

Bongers, H. 1947. Historical and Principles of Vocabulary Control. Wocopi, Worden.

Brown, R. 1973. A first language: the early states. Harvard University press, Cambridge.

Brown et al. (Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin). 1990.A Statistical Approach to Machine Translation, Computational Linguistics, 16(2), pages 79–85, June 1990.

Brochure on 'Language Technology Products' of the Resource Centre for Indian Language Technology Solutions-Tamil, Chennai.

Brown et al. (Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer). 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics, 19(2), pages 263–311, June 1993.

Chellamuthu, K.C. et al. Tamil University Machine Translation System (TUMTS), Thanjavur: Tamil University.

Bourbeau, L.(ed.). 1981. Linguistic documentation of computerized translation chain of TAUM-Aviation system. University of Montreal, May 1981 (I-VI) pp 77.

Chellamuthu, K.C. 2002. 'Russian to Tamil Machine Translation System at Tamil University,' in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, 74-83.

Darbari, H. 1999. Computer-assisted translation system – an Indian perspective. Machine Translation Summit VII, 13th-17th September 1999, Kent Ridge Digital Labs, Singapore. In Proceedings of MT Summit VII : MT in the Great Translation Era. pp.80-85.

Dash N.S. 2005. Corpus linguistics and Language Technology. Mittal Publications, New Delhi.

Dirix, P., Schuurman, I., and Vandeghinste, V. (2005). Metis II: Example-based machine translation using monolingual corpora - system description. In Proceedings of the 2nd Workshop on Example-Based Machine Translation, pages 43–50, Phuket, Tailand.

Durai Pandi. 2002. “English-Tamil Machine Translation System”, in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, page 86.

Fries C. 1940. American English Grammar. Appleton-Century-Crofts, New York.

Fries C and Traver A 1940. English word lists. A study of their adaptability and instruction. American Council of Education, Washington, DC.

Gale W.A. and Church K.W. 1993. “A program for aligning sentences in bilingual corpora”, Computational linguistics. 19(1):75-102.

Harshawardhan R., Augustine M.S., Soman K. P.. 2011. “Phrase based English-Tamil Translation System by Concept Labeling using Translation Memory”, in *Int. Journal of Computer Applications (IJCA)*, ISSN: 0975 – 8887, Vol. 20, no. 3, April, 2011.

Harshawardhan, R, Augustine M S. and Soman K. P.2011. “A Simplified Approach to Word Alignment Algorithm for English-Tamil Translation”, in *Indian Journal of Computer Science and Engineering (IJCSE)*, ISSN: 0976-5166, Vol. 2, No. 1, 2011.

Hiemstra, Djoerd. 1998. Multilingual domain modeling in twenty-one: automatic creation of a bi-directional lexicon from a parallel corpus. Technical report, University of Twente, Parlevink Group.

Hiemstra, Djoerd. August 1996. Using statistical methods to create a bilingual dictionary. Master’s thesis, Department of Computer Science, University of Twente.

Hutchins, W.J. 1986. *Machine translation: past, present, future*. Chichester (UK): Ellis Horwood; New York: Wiley.

Hutchins W J 1994 Research methods and system designs in machine translation: a ten-year review, 1984-1994. In: Machine Translation, Ten Years On, 12-14 November 1994, Cranfield University. 16pp.

Hutchins, W.J. 2005. The history of machine translation in a nutshell. http://www.hutchinsweb.me.uk/Nutshell-2005.pdf.

Hutchins, W. J. and Somers, H. L. 1992. An Introduction to Machine Translation. Academic Press, London, UK.

Ingram,D. 1978. ‘Sensori-motor development and language acquisition’, in Lock 1978, pp. 261-290.

Isabelle, P, Bourbeau L, Chevalier M, and Lepage S. 1978 TAUM-AVIATION: description d'un systrme de traduction automatisre de manuels d'entretien en arronautique. COL1NG-78, Bergen, Norway

Isabelle P and Bourbeau L. 1985. Computational Linguistics, Volume 11, Number 1, January-March 1985

Jain R, Sinha R.M.K., and Jain A. 2001. ANUBHARTI Using Hybrid Example-Based Approach for Machine Translation. In proceedings of Symposium on Translation Support Systems (SYSTRAN2001), February 15-17,2001. Kanpur. pp.123-130.

Jurafsky, D and Matin, J.H. 2000. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. New Jersey: Prentice Hall.

Kamakshi, S. and Rajendran S. 2004. Preliminaries to the preparation of a Machine Translation Aid to Translate Linguistics Texts written in English to Tamil. DLA publications, Thiruvananthapuram.

Kay M and Roscheisen M. 1993. Text-Translation Alignment, Computational Linguistics, 19(1), pp. 121–142, 1993.

Kennedy, G. 1992. 'Preferred ways of putting things', in Svartvik 1992, pp. 335-373.

King, M. (ed.) 1987. Machine translation today: the state of the art. Edinburgh University Press, Edinburgh.

Knight, K. 1999. A statistical machine translation tutorial workbook. http://www. isi.edu/natural-language/mt/wkbk.rtf. 35 pages.

Koehn, P. and Hoang, H. 2007. Factored translation models. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 868–876, Prague, Czech Republic.

Labov, W. 1969.'The logic of non-standard English', Georgetown Monographs on Language and Linguistics 22.

Lehmann, T. 1993. A grammar of modern Tamil. Pondicherry Institute of Linguistics and Culture, Pondicherry.

Leech, G 1991. 'A State of Art in Corpus Linguistics', in Aijmer and Altenberg 1991, pp. 8-29.

Leech, G. 1992. 'Corpora and theories of linguistic performance', in Svartvik 1992, pp 105-122.

Leech, G. 1993. 'Corpus annotation schemes', Literary and linguistic computing 8(4): 275-281.

Lopez, A. 2008. Statistical machine translation. ACM Computing Surveys, 40(3):1–49.

Manning C.D. and Schutze H. 2000. Foundations of Statistical Natural Language Processing, The MIT Press, 2000.

McEnery T. And Wilson A 1996. Corpus Linguistics. Edinburgh University Press, Edinburgh.

Melamed, I. Dan. 1999. Bitext maps and alignment via pattern recognition. Computational Linguistics, 25(1):107–130.

Melamed, I. Dan. 2000. Models of translational equivalence among words. Computational Linguistics, 26(2):221–249.

Nancy I and Veronis J. 1998. Word Sense Disambiguation: The State of the Art Computational Linguistics, 24(1), 1998.

Naskar S. and Bandyopadhyay Sivaji. 2005. Use of Machine Translation in India: Current Status. In proceedings of MT SUMMIT X, September 13-15, Phuket, Thailand. pp. 465-470.

Ney, H. 2005. “One decade of statistical machine translation.” In: AMTA (2005), i-12-17.

Och, F. J. 2005. Statistical machine translation: Foundations and recent advances. Tutorial at MT Summit X (http://www.mt-archive.info/MTS-2005-Och.pdf), Phuket, Thailand.

Och F.J. and Ney H. 2001. A Comparison of Alignment Models for Statistical Machine Translation, Proceedings of the 17th Conference on Computational Linguistics, pages 1086–1090, 2000.

Nirenburg, S. (ed.) 1987. Machine translation: theoretical and methodological issues. Cambridge University press, Cambridge.

University Press.

Nirenburg, S., Somers, H., & Wilks, Y. (eds.) Readings in machine translation. Cambridge, Mass.: MIT Press.

Palmer, H. 1933. Second interim report on English collocations. Institute for Research in English Teaching, Tokyo.

Quirk, R.1960. ‘Towards a description of English usage’, Transactions of Philosophical Society, pp. 4061.

Quirk, R., Greenbaum, S. 1988. A University Grammar of English, London: ELBS, Longman.

Quirk, R., Greenbaum, S, Leech, G.N. and Svartvik, J. 1972. A Grammar of Contemporary English. London: Longman.

Quirk, R., Greenbaum, S., Leech, G.N. and Svartvik, J. 1985. A Comprehensive Grammar of the English Language. London: Longman.

Rajendran, S. 2006. "Shallow Parsing in Tamil: the state of art", Language in India 6:7, www.languageinindia.com

Rajendran, S. 2006. "Languge Technology in Tamil", Language in India 6:8, www.languageinindia.com

Rajendran, S. et al. 2003. "Computational Morphology of Verbal Complex." In: B. Ramakrishna Reddy (edited) Word Structure in Dravidian, Kuppam: Dravidian University, & Language in India 3:4, www.languageinindia.com, April 2003.

Rajendran, S. and Kamakshi, S. Preliminaries to the Preparation of a Machine Aid to Translate Linguistic Texts in English into Tamil. Paper presented in Dravidian Linguists Conference.

Ramanathan A. Statistical Machine Translation, Ph.D. Seminar Report. Department of Computer Science and Engineering Indian Institute of Technology, Bombay aMumbai

Rangan, K. 1972. A Contrastive Analysis of the Grammatical Structures of Tamil and English. Unpublished Ph.D. Dissertation. Delhi: University of Delhi.

Rao D. 2001. Machine Translation in India: A Brief Survey. In proceedings of SCALLA2001 Conference, November 21-23, NCST, Bangalore, India. [Internet Source: http://elda.org/en/proj/scalla/SCALLA2001/SCALLA2001Rao.pdf].

Rekha, R U Anand kumar M, Dhanalakshmi.V , Soman K P, Rajendran S, 2010 "Morphological generator for Tamil a new data driven approach", Tamil Internet Conference 2010, June 2010, Cemmozhi maanaadu, Coimbatore, India.

Renuga Devi, V. 1997. Grammatical comparison of Tamil and English: A Typological Study. Madirai: Devi Publications.

Roberts A.H. Zarechnak, 1994. "Machine Translation", Currents Trends in Linguistics12 pp 2825-2870.

Saravanan S, Menon A.G. and Soman K.P. 2010. "Pattern Based English-Tamil Machine Translation", in Proceedings of Tamil Conference, Coimbatore, 2010.

Sereda, S.P. 1982. "Practical experience of machine translation", in Practical experience of machine translation. Proceedings of a conference, London 5-6 November 1981. Ed. V. Lawson, 119-123, North Holland, Amsterdam.

Sinha R. M. K., Jain R., and Jain A. 2001. Translation from English to Indian languages: ANGLABHARTI Approach. In proceedings of Symposium on Translation Support System STRANS 2001. February 15-17, IIT Kanpur, India. pp.167-172.

Computational Linguistics, Volume 11, Number 1, January-March 1985 “A survey of machine translation: Its history, current status, And future prospects,” Computational Linguistics, Volume 11, Number 1, January-March 1985.

Simoes, A.M.B. 2004. Parallel corpora word alignment and applications. Departamento de Inform´atica, Escola de Engenharia, Universidade do Minho, Braga, 2004

Slocum, J. (ed.) 1988. Machine translation systems. Cambridge: Cambridge University Press.

Somers H.L. 1999. “Example-based Machine Translation”, Machine Translation, 14, pages 113–157, 1999.

Souter, C. and Atwell, E. (Eds.) 1993, Corpus based computational Linguistics. Amsterdam: Rodopi.

Sperberg-McQueen C.M. and Burnard, L. 1994. Guidelines for electronic text encoding and interchange (P3). Text Encoding initiative, Chicago and Oxford.

Svarrvik, J. 1966. On voice in English verb. Mouton, The Hague.

Starvik, J. Ed. 1992. Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 – Stockholm, 4-8 August 1991. Berling, New York, Mouton De Gruyer. Thomas, J. and Short, M. (Eds.) 1996. Using Corpora for Language Research: Studies in the Honour of Geoffery Leech. London and New York: Addison Welsely Longman.

Thorndike, E. 1921. A teacher’s wordbook. Columbia Teachers College, New York.

Tognini–Boneli E. 2001. Corpus Linguistics at work. Amsterdam: John Benjammins.

Weaver, W. 1949. Translation, Machine Translation of Languages: Fourteen Essays, William Locke and Donald Booth (eds), pages 15–23, 1955.

Yamada K and Knight K. 2001. A Syntax-based Statistical Translation Model, Proceedings of the Conference of the Association for Computational Linguistics (ACL), 2001.